बुनियादी सांख्यिकीय तकनीक और आनुवंशिकी में इसका अनुप्रयोग विषय पर हिंदी कार्यशाला

(3 अगस्त 2022 – 5 अगस्त 2022)

Hindi Workshop on Basic Statistical Techniques and Its Application in Genetics

(3 August, 2022 – 5 August, 2022)

कार्यशाला समन्वयक

Workshop Coordinators

रंजित कुमार पॉल **Ranjit Kumar Paul**

प्रकाश कुमार **Prakash Kumar**

मो. यासीन **Md Yeasin**

सन्दर्भ पुस्तिका

Reference Manual

सांख्यिकीय आनुवंशिकी प्रभाग
भा.कृ.अनु.प.-भारतीय कृषि सांख्यिकी अनुसंधान संस्थान
लाइब्रेरी एवेन्यू, नई दिल्ली-110012

Division of Statistical Genetics
I.C.A.R.-Indian Agricultural Statistics Research Institute
Library Avenue, New Delhi - 110012

प्राक्कथन

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, सांख्यिकीय विज्ञान (सांख्यिकी, संगणक अनुप्रयोग एवं जैव सूचना विज्ञान) में प्रासंगिक कार्यों में संलग्न एक प्रमुख संस्थान है और कृषि अनुसंधान की गुणवत्ता को समृद्ध करने और नीतिगत निर्णय लेने के लिए कृषि विज्ञान में इनके विवेकपूर्ण संलयन में इसका प्रमुख योगदान है। 1930 में, तत्कालीन इंपीरियल काउंसिल ऑफ एग्रीकल्चरल रिसर्च (ICAR) के एक छोटे सांख्यिकीय अनुभाग के रूप में आरम्भ हो कर, संस्थान का कद ऊँचा उठा और राष्ट्रीय और अन्तर्राष्ट्रीय स्तर पर अपनी उपस्थिति दर्ज कराने में सक्षम हुआ है। संस्थान बहुत सक्रिय रूप से परामर्शदात्री सेवा प्रदान कर रहा है जिसने संस्थान को राष्ट्रीय कृषि अनुसंधान और शिक्षा प्रणाली (NARES) एवं राष्ट्रीय कृषि सांख्यिकी प्रणाली (NASS) दोनों में उपयोगिता निश्चित कराने में सफलता हासिल की है। संस्थान ने NARES में एक उच्च स्तरीय सांख्यिकीय संगणना के लिए उपयुक्त वातावरण बनाने में अग्रणी भूमिका निभाई है।

अनुसंधान कार्यों से लिए गए सांख्यिकीय रूप से वैध और सार्थक निष्कर्ष ही गुणवत्तापूर्ण शोध की नींव बनाते हैं और नीति निर्धारण एवं कार्यक्रम क्रियान्वयन में एक महत्त्वपूर्ण भूमिका निभाते है। इसलिए यह आवश्यक है कि आँकड़ों के संकलन एवं विश्लेषण के लिए ठोस सांख्यिकीय पद्धति अपनाई जाए। संस्थान द्वारा आयोजित कार्यशाला, कृषि विज्ञान के अनुसंधान एवं योजना निर्धारण में कार्यरत प्रयोक्ताओं के लिए सांख्यिकीय तकनीकों में प्रगति का मूल्यांकन करने में बहुत उपयोगी सिद्ध हुए हैं।

वैज्ञानिकों एवं तकनीकी अधिकारियों के लिए **बुनियादी सांख्यिकीय तकनीक और आनुवंशिकी में इसका अनुप्रयोग** पर हिन्दी कार्यशाला की संकल्पना, संस्थान के हिन्दी के प्रचार एवं प्रसार की दिशा में एक कदम है। मुझे विश्वास है कि इस कार्यशाला के दौरान ग्रहण किया गया ज्ञान सहभागियों को सांख्यिकीय तकनीकों को बेहतर समझ रखने में सक्षम बनाने के साथ–साथ, उपयुक्त और आधुनिक सांख्यिकीय पद्धतियों द्वारा आँकड़ों का विश्लेषण करने में भी लाभदायक होगा।

पाठ्यक्रम सामग्री अत्यधिक अनुप्रयोग उन्मुख है। इस कार्यशाला में शामिल संकाय सदस्य कृषि सांख्यिकी के क्षेत्र में प्रख्यात सांख्यिकीविद हैं। कार्यशाला के लिए ई–संदर्भ संहिता तैयार की गई है और इस ई–संदर्भ संहिता में दिये गए व्याख्यान, विषय का विस्तृत विवरण प्रदान करते हैं। मुझे आशा है कि सहभागियों के लिए यह संदर्भ संहिता अत्यन्त उपयोगी होगी। डॉ. अजीत, प्रभागाध्यक्ष (का.), सांख्यिकीय आनुवंशिकी प्रभाग, डॉ. रंजित कुमार पॉल, प्रकाश कुमार एवं डॉ. मो. यासीन इस ऑनलाईन कार्यशाला के समन्वयक और अन्य सहयोगी इस संदर्भ संहिता को समय से निकालने के लिए सराहना के पात्र हैं।

2/8/22

नई दिल्ली
अगस्त 02, 2022

(राजेन्द्र प्रसाद)
निदेशक
भा.कृ.अनु.प.–भा.कृ.सां.अ.सं.

आमुख

भा.कृ.अनु.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, कृषि सांख्यिकी एवं सूचना विज्ञान विषय पर एक अग्रणी संस्थान है। यह संस्थान भारतीय कृषि अनुसंधान परिषद् के कृषि शिक्षा प्रभाग के मानव संसाधन विकास कार्यक्रम के तत्वावधान में कृषि सांख्यिकी एवं संगणक अनुप्रयोग में उच्च संकाय प्रशिक्षण केन्द्र, जो कि पहले उच्च अध्ययन केन्द्र के नाम से जाना जाता था, के रूप में भी कार्य कर रहा है। उच्च संकाय प्रशिक्षण केन्द्र के अन्तर्गत आयोजित किये जाने वाले प्रशिक्षण कार्यक्रमों के अतिरिक्त, संस्थान भारतीय कृषि अनुसंधान परिषद् के कृषि शिक्षा विभाग द्वारा प्रायोजित ग्रीष्मकालीन/शीतकालीन स्कूल तथा विभिन्न राष्ट्रीय एवं अर्न्तराष्ट्रीय संस्थानों की आवश्यकतानुसार अन्य प्रशिक्षण कार्यक्रम भी आयोजित करता है। वर्तमान हिन्दी कार्यशाला संस्थान के हिन्दी गतिविधियों के सुचारु संचालन हेतु हिन्दी एकक के निदेशानुसार आयोजित किया गया है।

अध्ययन एवं अध्यापन के सभी विषयों में सांख्यिकी की भूमिका महत्त्वपूर्ण है। यह अत्यन्त आवश्यक है कि संकलित आँकड़े उपयुक्त एवं परिशुद्ध हों। अनुसंधान की गुणवत्ता का उच्चस्तर बनाये रखने एवं बढ़ाने के लिए यह अत्यन्त महत्त्वपूर्ण एवं आवश्यक है कि आँकड़ों के संकलन, विश्लेषण एवं परिणामों की व्याख्या के लिए उचित सांख्यिकीय पद्धतियों को अपनाया जाये। शोध कार्यक्रमों के सांख्यिकी रूप से वैध निष्कर्ष नीति निर्धारण विशेष रूप से विकास कार्यक्रमों एवं कार्यक्रम क्रियान्वयन में महत्त्वपूर्ण भूमिका निभाते है। सही योजना एवं कार्यक्रम क्रियान्वयन के लिए आँकड़ों का सामयिक एवं परिशुद्ध होना अनिवार्य है। राष्ट्रीय महत्व के विषयों से संबंधित आँकड़ों के संकलन एवं विश्लेषण की पद्धतियाँ विकसित करने में संस्थान अग्रणीय है। राष्ट्रीय कृषि साख्यिकी प्रणाली के अन्तर्गत विभिन्न प्राचलों के आकलन संबंधी पद्धतियाँ विकसित करने के लिए संस्थान का मूलभूत योगदान है। संस्थान द्वारा आयोजित प्रशिक्षण कार्यक्रम इन सभी विषयों पर प्रकाश डालते हैं तथा अनुसंधान एवं नीति निर्धारण में कार्यरत प्रयोक्ताओं को सांख्यिकी तकनीकों में विकास संबंधी जानकारी देने में अत्यन्त उपयोगी है।

बुनियादी सांख्यिकीय तकनीक और आनुवंशिकी में इसका अनुप्रयोग पर इस कार्यशाला का उद्देश्य सहभागियों को आँकड़ों के सांख्यिकीय विश्लेषण की परिष्कृत सांख्यिकीय तकनीकों में विकास एवं सॉफटवेयर के प्रयोग से अवगत कराना है। परिणामों की व्याख्या करने एवं प्रस्तुतिकरण पर विशेष बल दिया जायेगा। हिन्दी कार्यशाला को इस प्रकार तैयार किया गया है कि यह सिद्धान्त एवं अनुप्रयोग का मिला–जुला रूप है।

हम सभी संकाय सदस्यों का हार्दिक धन्यवाद करते हैं जिन्होंने इस हिन्दी कार्यशाला को सार्थक एवं सफल बनाने में अपना अमूल्य समय दिया है। इस ई–संदर्भ संहिता को तैयार करने में त्रुटियों को कम करने का हर सम्भव प्रयास किया गया है, फिर भी कमियाँ हो सकती है। भा.कृ.अ.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान या लेखक, इस ई–संदभ संहिता में दी गई सामग्री के प्रयोग द्वारा हुई किसी भी प्रकार की हानि के लिए उत्तरदायी नहीं होंगे। हम श्री उदयवीर सिंह एवं हिन्दी एकक के आभारी है जिन्होंने हमें यह कार्यशाला आयोजित करने का अवसर दिया। हम डॉ. राजेंद्र प्रसाद, निदेशक, भा.कृ.अ.प.–भारतीय कृषि सांख्यिकी अनुसंधान संस्थान एवं डॉ. अजीत, प्रभागाध्यक्ष, सांख्यिकीय आनुवंशिकी प्रभाग, के आभारी हैं, जिन्होंने हमारा मार्गदर्शन किया और इस पाठ्यक्रम में सतत् रूचि बनाए रखी व हमें सभी जरूरी सुविधाएँ उपलब्ध कराई। हम उन सबके प्रति भी आभारी हैं जिन्होंने अपने अथक प्रयासों से इस संदर्भ संहिता को तैयार करने में मदद की है। इस हिन्दी कार्यशाला सामग्री को और अधिक उपयोगी बनाने के लिए आपके बहुमूल्य सुझाव आमंत्रित हैं।

नई दिल्ली अगस्त 02, 2022	(रंजित कुमार पॉल) कार्यशाला समन्वयक	(प्रकाश कुमार) कार्यशाला सह समन्वयक	(मो. यासीन) कार्यशाला सह समन्वयक	(अजीत) कार्यशाला सलाहकार

बुनियादी सांख्यिकीय तकनीक और आनुवंशिकी में इसका अनुप्रयोग

व्याख्यान सारणी

हिन्दी फोंट्स और यूनिकोड का उपयोग करके हिन्दी लेखन

उदय वीर सिंह
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
hindi.iasri@icar.gov.in

यूनिकोड क्या है?

सर्वप्रथम यह समझना आवश्यक है कि यूनिकोड क्या है ? क्या यूनिकोड कोई फोंट है? क्या यूनिकोड कोई टंकण टूल है? कम्प्युटर में टाइप करने के लिए अनेक फोंट्स उपलब्ध होते हैं। सभी फोंट्स को दो भागों में बांटा गया है,

1. नॉन-यूनिकोड 2. यूनिकोड

हिन्दी में टाइप करने के लिए प्रायः हम लोग क्रुतिदेव अथवा मंगल फोंट्स का प्रयोग करते हैं। इनमें कृतिदेव फॉन्ट नॉन-यूनिकोड के अंतर्गत आता है जबकि मंगल फॉन्ट यूनिकोड के अंतर्गत आता है। यूनिकोड एक वैज्ञानिक तकनीक है, एक व्यवस्था है, जो प्रचलित प्रत्येक लिपि के वर्णमाला के अक्षर को चार अंकों का विशेष कोड या नम्बर प्रदान करता है इन्टरनेट पर जो हम हिन्दी में लिखे हुए लेख देखते हैं, यह सब यूनिकोड तकनीक का इस्तेमाल करके ही लिखे जाते हैं।

नॉन-यूनिकोड और यूनिकोड में अंतर :

नॉन-यूनिकोड :

जब हम नॉन-यूनिकोड फोंट्स अर्थात क्रुतिदेव फॉन्ट का उपयोग करके हिन्दी में टाइप करते हैं एवं टाइप किये गये पत्र, लेख अथवा संदेश को इंटरनेट या किसी और माध्यम से एक कम्प्युटर से दूसरे कम्प्युटर पर अथवा एक उपयोगकर्ता से दूसरे उपयोगकर्ता को भेजते हैं तो अक्सर यह होता है कि यदि दूसरे उपयोगकर्ता के पास कम्प्युटर में क्रुतिदेव फॉन्ट उपलब्ध नहीं है तो उस कम्प्युटर पर भेजा हुआ पत्र या लेख पढ़ने में नहीं आता है, उसका स्वरूप बदल जाता है वह अपाठ्य हो जाता है, चाहे दूसरे उपयोगकर्ता के पास क्रुतिदेव फॉन्ट उपलब्ध हो भी, तब भी हिन्दी वर्णमाला के कुछ अक्षरों मे अपभ्रंश हो जाता है, उनका स्वरूप ही बदल जाता है अर्थात उस वाक्य का अर्थ ही बदल जाता है।

हम सभी इंग्लिश में तो टाइप करने के अभ्यस्त हैं। क्योंकि हमें कम्प्युटर में इंग्लिश के कीबोर्ड पर इंग्लिश के अक्षरों की तो जानकारी होती है, लेकिन इंग्लिश के कीबोर्ड पर हिन्दी भाषा/ लिपि की वर्णमाला के अक्षरों की जानकारी नहीं होती है, कि हिन्दी का कौन सा अक्षर कहाँ पर अत: नॉन-यूनिकोड फॉन्ट का उपयोग केवल हिन्दी टाइपिस्ट अथवा टाइप करने का अभ्यास करने वाले ही टाइप कर सकते हैं।

यूनिकोड:

जब हम यूनिकोड फॉन्ट का उपयोग करके टाइप किये गये पाठ, सामग्री, पत्र, लेख अथवा संदेश को इंटरनेट या किसी और माध्यम से एक कम्प्युटर से दूसरे कम्प्युटर पर अथवा एक उपयोगकर्ता से दूसरे उपयोगकर्ता को भेजते हैं तो उसका स्वरूप बिलकुल भी नहीं बदलता है, वह पूरी दुनिया में कही भी पढ़ा जा सकता है, इसके लिये किसी विशेष फॉन्ट की आवश्यकता नहीं होती है। यूनिकोड फॉन्ट का उपयोग कम्प्युटर में अँग्रेजी के कीबोर्ड पर अँग्रेजी के अक्षरों की जानकारी रखने वाले भी हिन्दी में बहुत अच्छी तरह से टाइप कर सकते है।

यूनिकोड के लाभ :

इसका लाभ यह है कि चाहे किसी भी सॉफ्टवेयर में या किसी भी भाषा में यूनिकोड का प्रयोग किया जा सकता है। यूनिकोड में टाइप किये गये पाठ या सामग्री को कहीं भी ले जाने पर उसका स्वरूप नहीं बदलता है, वह पूरी दुनिया में कही भी पढी जा सकती है, इसके लिये किसी विशेष फान्ट की आवश्यकता नहीं होती है।

यूनिकोड विश्व की ज्यादातर भाषाओं में बदला जा सकता है। यूनिकोड मानक को एपल, एच.पी., आई.बी.एम., माइक्रोसौफ्ट, औरेकल जैसी प्रमुख कम्पनियो ने अपनाया है । यूनिकोड में हिन्दी तथा अन्य भाषाओ में कम्प्युटर पर अँग्रेजी की तरह आसानी से 100% कार्य किया जा सकता है। जैसे- वर्ड प्रोसेसिंग, डाटा प्रोसेसिंग, ई-मेल, वैबसाइट निर्माण आदि किए जा सकते हैं। यूनिकोड तीन प्रकार का होता है:

1. यूटीएफ़-8 2. यूटीएफ़-16 3. यूटीएफ़-32

भारतीय भाषाओ के लिए यूनिकोड एड्कोडिंग के लिए यूटीएफ़-8 का प्रयोग किया जाता है।

यूनिकोड के अन्य लाभ :

1. आप बिना हिन्दी टाइप जाने हिन्दी में टाइप कर सकते हैं ।
2. आप गूगल सर्च में हिन्दी में सर्च कर सकते हैं ।

3. हिन्दी में ई-मेल भेज सकते हो।
4. कम्प्यूटर में विभिन्न फाइल और फोल्डरों के नाम हिन्दी में रख सकते हैं ।
5. हिन्दी में चैट कर सकते हैं ।
6. हिन्दी में वेब साइट या ब्लाग बना सकते हैं ।
7. वर्ड और एक्सेल में बिना हिन्दी फॉन्ट डाउनलोड किये हिन्दी में टाइपिंग की जा सकती है।
8. फेसबुक और ट्विटर जैसे सोशल नेटवर्किंग साइट पर आसानी से हिन्दी में लिखा जा सकता है।
9. यूनिकोड में लिखी किसी भी सामग्री को आसानी से दूसरी भाषा में परिवर्तित किया जा सकता है।

सरकारी कार्यालयों में यूनिकोड का प्रयोग:

भारत सरकार द्वारा यूनिकोड को अपने सभी कार्यालयों में अनिवार्य कर दिया गया है, भारत सरकार की सभी साइटों पर यूनिकोड का प्रयोग किया जा रहा है। यहॉ तक कि नई भर्तियों के लिये अभ्यर्थियों को यूनिकोड फॉन्ट में टाइपिंग टेस्ट भी अनिवार्य कर दिया गया है।

यूनिकोड के अंतर्गत मंगल फॉन्ट का उपयोग कर टाइप करने के कुछ उदाहरण:

Bhartiya Krishi Sankhyiki Anusandhan Sansthan

भारतीय कृषि सांख्यिकी अनुसंधान संस्थान

Hindi Karyshala

हिन्दी कार्यशाला

Krishi mein pratidarsh taknikon ka anuprayog evam aankdon ka vishleshan

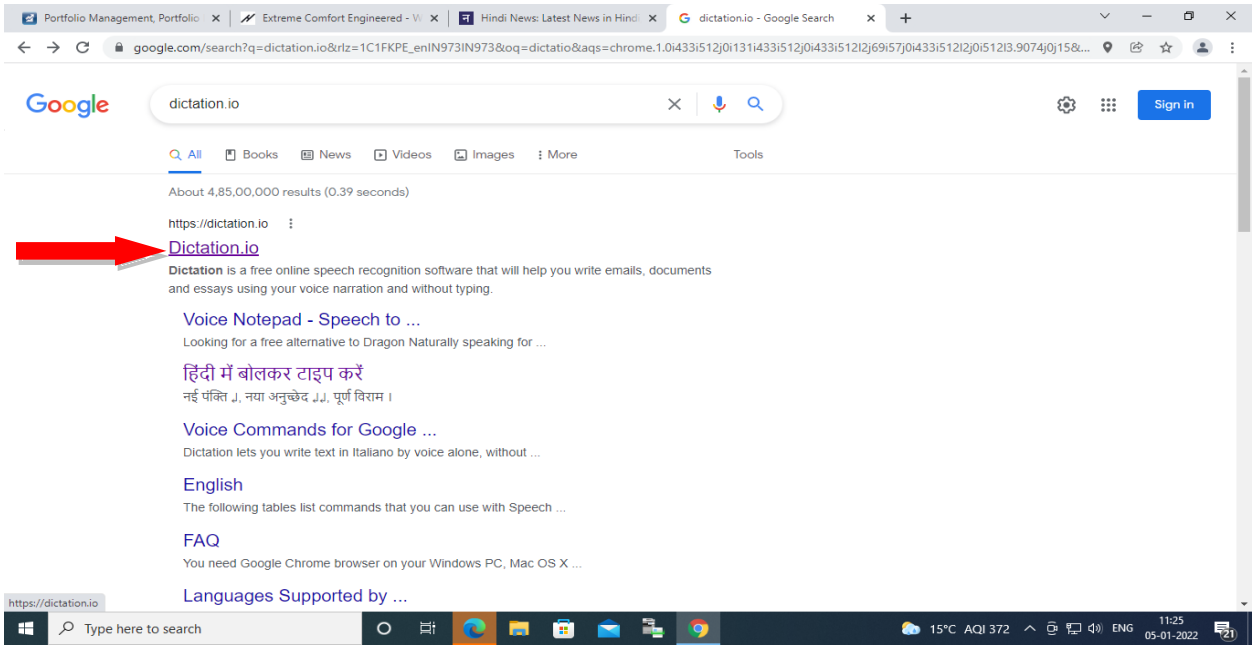
कृषि में प्रतिदर्श तकनीकों का अनुप्रयोग एवं आंकड़ों का विश्लेषण

Pramukh Vaigyanik	Takniki Adhikari	Prashsnik Adhikari
प्रमुख वैज्ञानिक	तकनीकी अधिकारी	प्रशासनिक अधिकारी

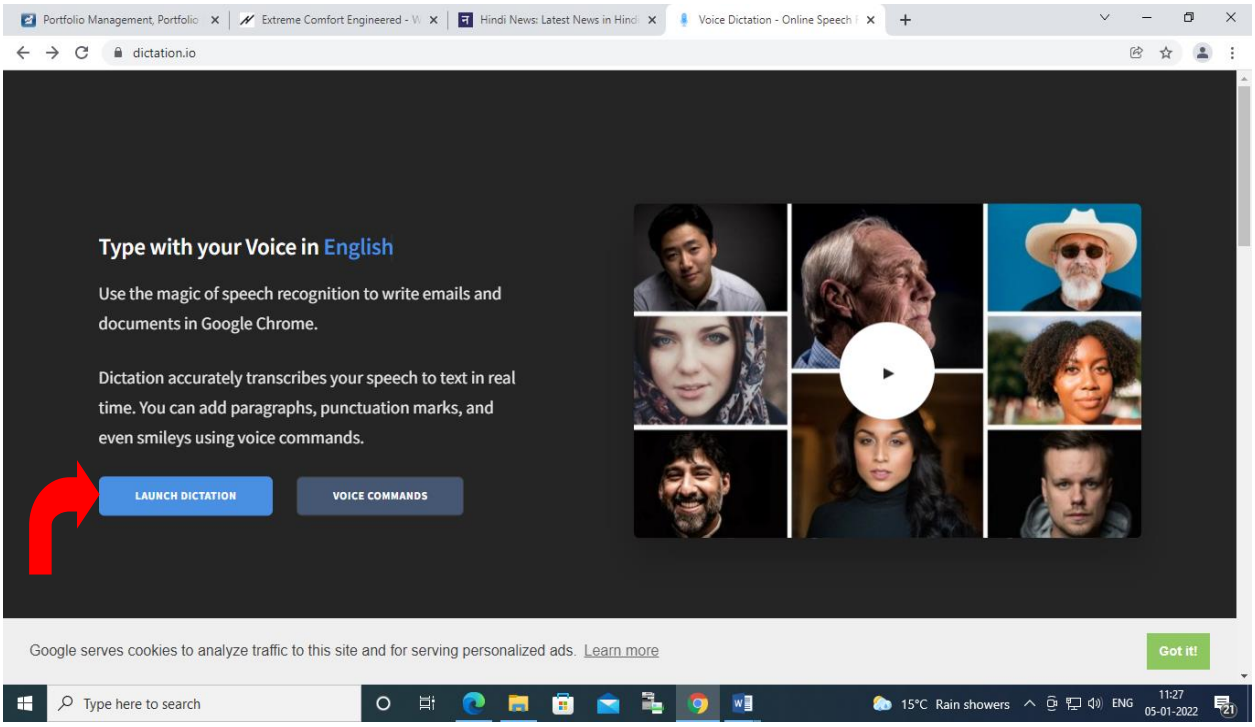
Ek Kisaan ki Mera Gaon mera gaurav pariyojna ke antargat krishi Vagyanik se Vartalap

एक किसान की मेरा गाँव मेरा गौरव परियोजना के अंतर्गत कृषि वैज्ञानिक से वार्तालाप

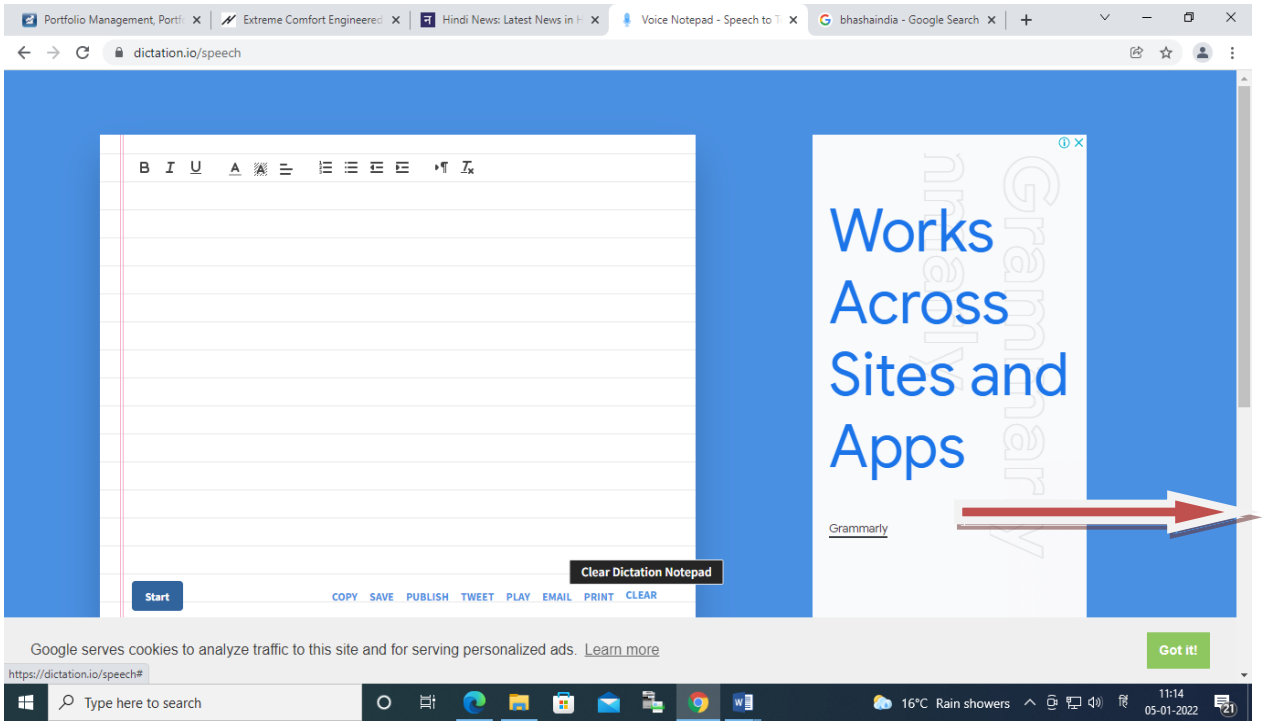
यूनिकोड के अतिरिक्त कम्प्युटर पर ऑनलाइन आवाज/ बोलने से टाइप/ टंकण करने की सुविधा



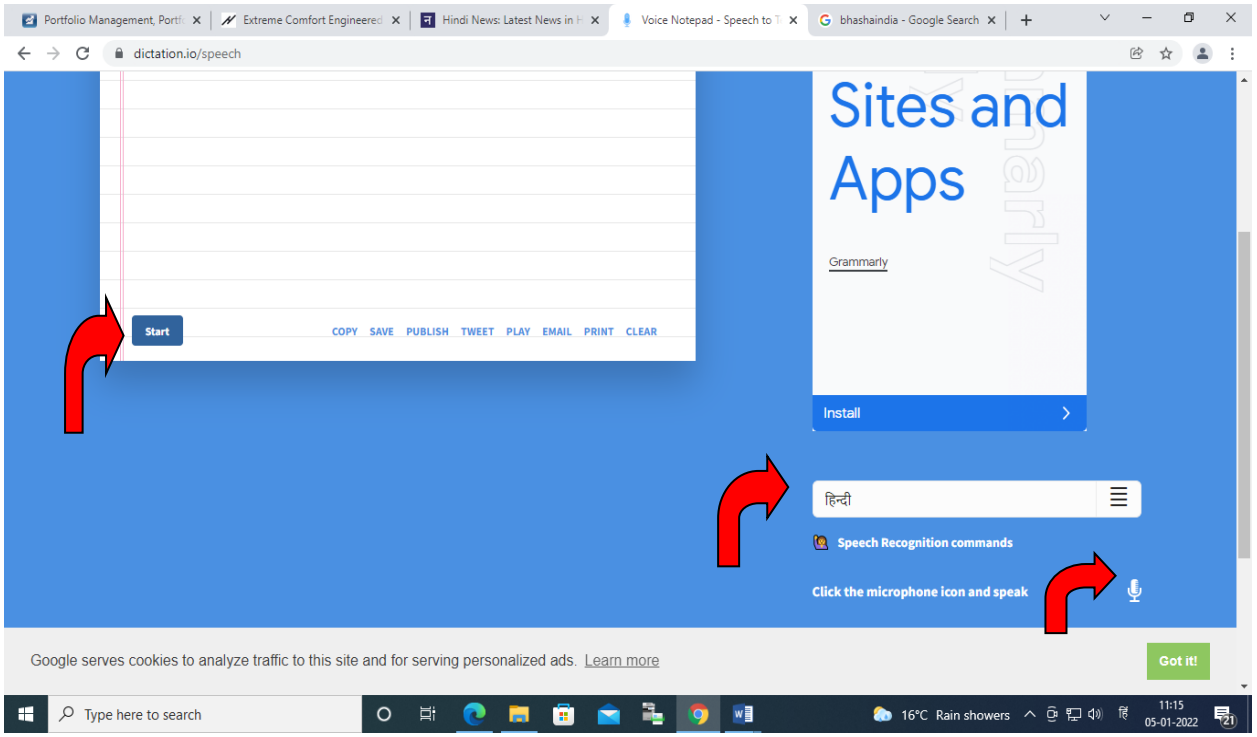
गूगल में dictation.io सर्च करके web site https://dictation.io पर क्लिक करें।



Launch Dictation पर क्लिक करें।

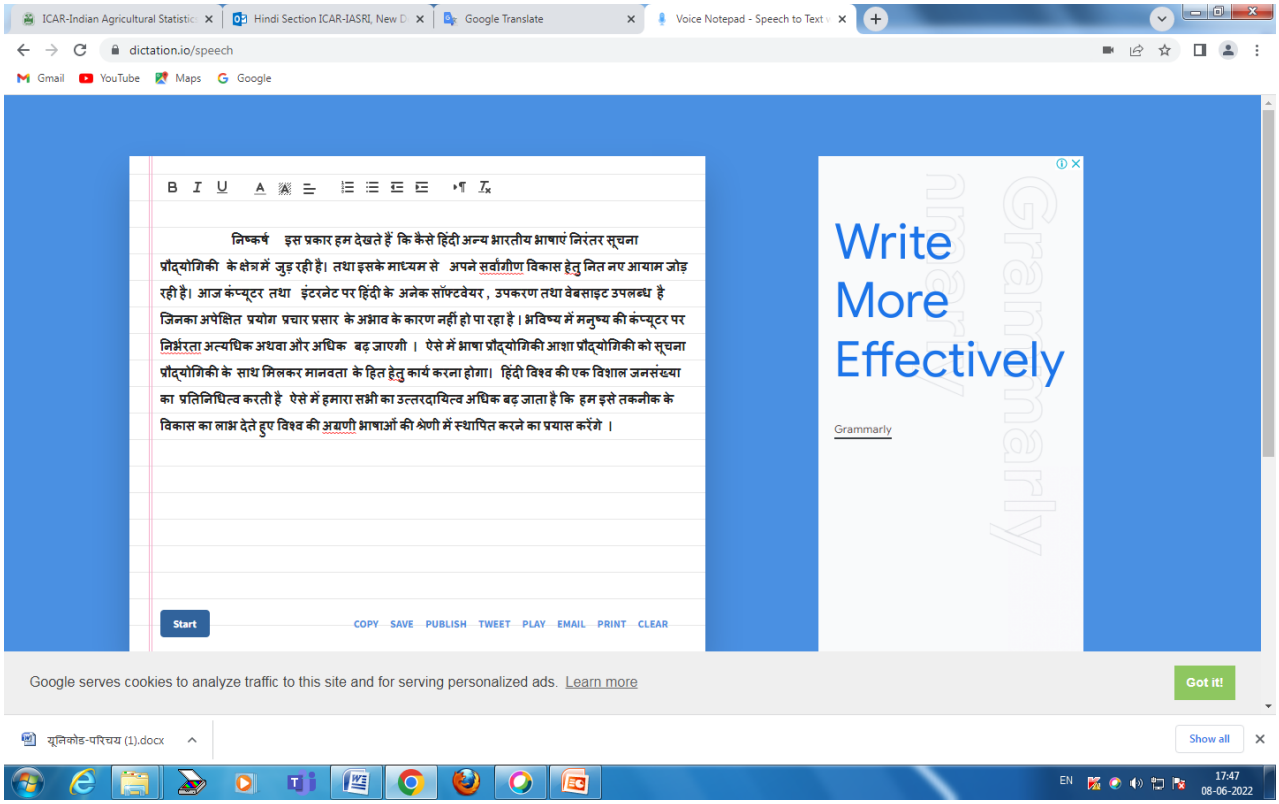


रूलर को स्क्रोल कर के नीचे लाएँ ।



स्क्रोल नीचे करने के बाद भाषा चुनें और माइक्रोफोन अथवा स्टार्ट पर क्लिक करें

विभिन्न चिन्हों को टाइप करने के लिए उक्त शब्दों का प्रयोग करें ।



ऑन लाइन टाइप करने के पश्चात कॉपी करके एम.एस. वर्ड में पेस्ट कर सकते हैं और एडिट भी कर सकते हैं।

ऑन लाइन टाइप करने के बाद एम.एस. वर्ड में कॉपी, पेस्ट और एडिट किया हुआ लेख

निष्कर्ष

इस प्रकार हम देखते हैं कि कैसे हिंदी एवं अन्य भारतीय भाषाएं निरंतर सूचना प्रौद्योगिकी के क्षेत्र में जुड़ रही है। तथा इसके माध्यम से अपने सर्वांगीण विकास हेतु नित नए आयाम जोड़ रही है । आज कंप्यूटर तथा इंटरनेट पर हिंदी के अनेक सॉफ्टवेयर, उपकरण तथा वेबसाइट उपलब्ध है जिनका अपेक्षित प्रयोग प्रचार प्रसार के अभाव के कारण नहीं हो पा रहा है । भविष्य में मनुष्य की कंप्यूटर पर निर्भरता और अधिक बढ़ जाएगी । ऐसे में भाषा प्रौद्योगिकी को सूचना प्रौद्योगिकी के साथ मिलकर मानवता के हित हेतु कार्य करना होगा। हिंदी विश्व की एक विशाल जनसंख्या का प्रतिनिधित्व करती है। ऐसे में हम सभी का उत्तरदायित्व और अधिक बढ़ जाता है कि हम इसे तकनीक के विकास का लाभ देते हुए विश्व की अग्रणी भाषाओं की श्रेणी में स्थापित करने का प्रयास करें ।

बुनियादी सांख्यिकीय तकनीक

डॉ. अजित
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
ajit@icar.gov.in

---

1. परिचय

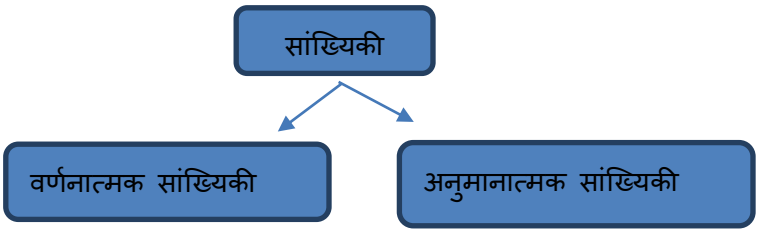
'सांख्यिकी' शब्द लैटिन शब्द 'स्टेटस' या इटेलियन शब्द 'स्टेटिस्टा' या जर्मन शब्द 'स्टेटिस्टिक' से लिया गया है, जिनमें से प्रत्येक का अर्थ 'पोलिटिकल स्टेट' है। सांख्यिकी एक व्यापक संकल्पना है जो विभिन्न क्षेत्रों में अनुप्रयोगों की विशेषता रखती है। सामान्य तौर पर, सांख्यिकी को डेटा को एकत्र करने, विश्लेषण करने, व्याख्या करने और निष्कर्ष निकालने की प्रक्रिया के रूप में परिभाषित किया जा सकता है। दूसरे शब्दों में, सांख्यिकी वैज्ञानिकों और गणितज्ञों द्वारा प्राप्त आंकड़ों से विश्लेषण और निष्कर्ष निकालने के लिए स्थापित दृष्टिकोण है। डेटा के संग्रह, प्रसंस्करण, व्याख्या और प्रस्तुति के साथ कुछ भी करने वाली हर चीज सांख्यिकी के दायरे में आती है।

सांख्यिकी की परिभाषा: सांख्यिकी गणित की एक शाखा है जो डेटा एकत्र करने, व्यवस्थित करने, सारांशित करने, प्रस्तुत करने और विश्लेषण करने के साथ-साथ वैध परिणाम प्रदान करने और उचित निर्णयों की व्याख्या करने से संबंधित है।

दूसरे शब्दों में, सांख्यिकीविद् निम्न लिए कार्यप्रणाली देते हैं

- डिजाइन: अनुसंधान परियोजनाओं की योजना बनाना और उनका संचालन करना।
- विवरण: डेटा सारांश और अन्वेषण।
- अनुमान: डेटा के बारे में भविष्यवाणियां और अनुमान लगाना।

सांख्यिकी को दो वर्गों में विभाजित किया जा सकता है; एक वर्णनात्मक सांख्यिकी है और दूसरा अनुमानात्मक सांख्यिकी है।



वर्णनात्मक आँकड़े सार्थक तरीके से डेटा का वर्णन करने, दिखाने या सारांशित करने में मदद करते हैं। वर्णनात्मक आँकड़े हमें डेटा को व्यवस्थित और सारांशित करने के लिए टूल, टेबल, ग्राफ़, औसत, रेंज, सहसंबंध प्रदान करते हैं। उदाहरण: मेसर ऑफ़ सेंट्रल टेन्डेन्सी, मेसर ऑफ़ डिस्पेर्सेन, स्क्यूनिस, करटोसिस आदि।

अनुमानात्मक आँकड़े नमूना मूल्यों को देखकर जनसंख्या के गुणों को समझने में मदद करते हैं। अनुमानात्मक आँकड़े एस्टिमेशन ऑफ़ पैरामीटर्स और टेस्टिंग ऑफ़ हाइपोथिसिस से संबंधित हैं।

इस खंड में हमने संक्षेप में वर्णनात्मक आंकड़ों पर चर्चा की जैसे केंद्रीय प्रवृत्ति के माप, फैलाव के उपाय, तिरछापन और कुर्टोसिस

2. केंद्रीय प्रवृत्ति का माप

केंद्रीय प्रवृत्ति एक सांख्यिकीय माप है जो एक एकल मान निर्धारित करता है जो वितरण के केंद्र का सटीक वर्णन करता है। केंद्रीय प्रवृत्ति का उद्देश्य एकल मूल्य की पहचान करना है जो डेटा के पूरे सेट के लिए सबसे अच्छा प्रतिनिधि है।

केंद्रीय प्रवृत्ति के विभिन्न माप हैं:

- माध्य
    - अंकगणित माध्य
    - गुणोत्तर माध्य
    - हरात्मक माध्य

- मध्यमान
- **मोड**
- चतुर्थकों
- दशमांश
- प्रतिशतता

माध्य (अंकगणित माध्य: A.M):

माध्य केंद्रीय प्रवृत्ति का सबसे अधिक इस्तेमाल किया जाने वाला माप है। माध्य डेटा की गणना के लिए अंतराल या अनुपात पैमाने पर मापा गया संख्यात्मक मान होना चाहिए। माध्य की गणना करने के लिए, हम डेटा सेट के अवलोकन को जोड़ते हैं और फिर अवलोकन की संख्या से विभाजित करते हैं।

$$\text{माध्य} = \frac{\text{सभी प्रेक्षणों का योग}}{\text{अवलोकन की कुल संख्या}}$$

2.1.1 **सरल माध्य**: मान लीजिए $X_1, X_2, \ldots, X_n$ एक डेटा सेट के n अवलोकन हैं। अंकगणित माध्य द्वारा दिया गया है

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

आवृत्ति वितरण के लिए माध्य: मान लीजिए $X_1, X_2, \ldots, X_n$ संगत आवृत्तियों के साथ अवलोकन हैं $f_1, f_2, \ldots, f_n$ तथा $\sum_{i=1}^{n} f_i = N$. अंकगणित माध्य द्वारा दिया गया है

$$\bar{X} = \frac{\sum_{i=1}^{n} f_i X_i}{N}$$

माध्य के गुण:

- यह उत्पत्ति के परिवर्तन के साथ-साथ पैमाने के परिवर्तन पर भी निर्भर करता है।

$$U = a + hX$$

Origin Scale

तब $\bar{U} = a + h\bar{X}$

- यदि $\bar{X}_1$ और $\bar{X}_2$ क्रमशः n_1 और n_2 टिप्पणियों के साथ मानों के दो सेटों के साधन हैं, तो उनका संयुक्त माध्य निम्न द्वारा दिया जाता है

$$\bar{X} = \frac{n_1\bar{X}_1 + n_2\bar{X}_2}{n_1 + n_2}$$

- मानों के समुच्चय के विचलनों का उनके माध्य से बीजगणितीय योग शून्य होता है।

$$\sum_{i=1}^{n} (X_i - \bar{X}) = 0$$

- इसके माध्य के सापेक्ष मानों के समुच्चय के विचलन के वर्गों का योग न्यूनतम है

$$\sum_{i=1}^{n} (X_i - A)^2 \text{ minimum when } A = \bar{X}$$

माध्य के गुण:

- समझने में आसान
- गणना करने में आसान।
- इसे सख्ती से परिभाषित किया गया है।
- यह सभी अवलोकनों पर आधारित है।
- यह नमूना उतार-चढ़ाव से कम से कम प्रभावित होता है।
- यह आगे गणितीय उपचार करने में सक्षम है।

माध्य के दोष:

- यह चरम मूल्यों से प्रभावित है।
- इसकी गणना ओपन एंड क्लास आवृत्ति वितरण के लिए नहीं की जा सकती है।

- इसे ग्राफिक रूप से स्थित नहीं किया जा सकता है।
- गुणात्मक विशेषता के लिए इसकी गणना नहीं की जा सकती है।
- यदि डेटा श्रृंखला में कोई अवलोकन गायब है तो इसकी गणना नहीं की जा सकती है।
- यह अत्यधिक विषम वितरण के लिए उपयुक्त नहीं है।

2.1.2 ज्यामितीय माध्य (जीएम):

n प्रेक्षणों के लिए, ज्यामितीय माध्य उनके गुणनफल का n^{th} मूल होता है।

गैर-आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \dots, X_n$ डेटा सेट के n अवलोकन हैं। ज्यामितीय माध्य के रूप में परिभाषित किया गया है

$$G = (X_1 * X_2 * \dots * X_n)^{1/n}$$

आवृत्ति वितरण के लिए: मान लें कि $X_1, X_2, \dots, X_n$ संगत आवृत्तियों के साथ अवलोकन हैं $f_1, f_2, \dots, f_n$ तथा $\sum_{i=1}^{n} f_i = N$. ज्यामितीय माध्य के रूप में परिभाषित किया गया है

$$G = ({X_1}^{f_1} * {X_2}^{f_2} * \dots * {X_n}^{f_n})^{1/N}$$

ज्यामितीय माध्य का उपयोग:

- औसत सापेक्ष परिवर्तन, औसत अनुपात और प्रतिशत मापें
- सूचकांक संख्या के निर्माण के लिए सर्वश्रेष्ठ औसत

ज्यामितीय माध्य के गुण:

- यह सभी अवलोकनों पर आधारित है।
- यह नमूने के उतार-चढ़ाव से प्रभावित नहीं होता है।
- यह आगे गणितीय उपचार करने में सक्षम है।

ज्यामितीय माध्य के दोष:

- यदि कोई मान शून्य है, तो इसकी गणना नहीं की जा सकती है।
- यह चरम मूल्यों से प्रभावित है।
- इसकी गणना ओपन एंड क्लास बारंबारता वितरण के लिए नहीं की जा सकती है।
- इसे ग्राफिक रूप से स्थित नहीं किया जा सकता है।
- गुणात्मक विशेषता के लिए इसकी गणना नहीं की जा सकती है।
- यदि डेटा श्रृंखला में कोई अवलोकन गायब है तो इसकी गणना नहीं की जा सकती है।

हार्मोनिक माध्य (H.M.):

हार्मोनिक माध्य समुच्चयों के प्रेक्षणों के व्युत्क्रमों के अंकगणितीय माध्य का व्युत्क्रम है।

गैर-आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \dots, X_n$ डेटा सेट के n अवलोकन हैं। हार्मोनिक माध्य को परिभाषित किया गया है:

$$H = \frac{n}{\sum_{i=1}^{n} 1/X_i}$$

आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \dots, X_n$ सं गत आवृत्तियों के साथ अवलोकन हैं $f_1, f_2, \dots, f_n$ तथा $\sum_{i=1}^{n} f_i = N$ हार्मोनिक माध्य को परिभाषित किया गया है:

$$H = \frac{N}{\sum_{i=1}^{n} f_i/X_i}$$

हार्मोनिक माध्य का उपयोग:

- उस परिवर्तन को मापें जहां एक चर के मूल्यों की तुलना दूसरे चर की स्थिर मात्रा के साथ की जाती है, जैसे समय, एक निश्चित समय के भीतर तय की गई दूरी, एक इकाई पर खरीदी या बेची गई मात्रा।

हार्मोनिक माध्य के गुण:

- यह छोटी वस्तु को अधिक भार देता है और बड़े मूल्यों को कम भार देता है।
- यह सभी अवलोकनों पर आधारित है।
- यह नमूने के उतार-चढ़ाव से प्रभावित नहीं होता है।
- यह आगे गणितीय उपचार करने में सक्षम है।

हार्मोनिक माध्य के अवगुण:

- यदि कोई मान शून्य है, तो इसकी गणना नहीं की जा सकती है।
- यह चरम मूल्यों से प्रभावित है।
- इसकी गणना ओपन एंड क्लास बारंबारता वितरण के लिए नहीं की जा सकती है।
- इसे ग्राफिक रूप से स्थित नहीं किया जा सकता है।
- गुणात्मक विशेषताओं के लिए इसकी गणना नहीं की जा सकती है।
- यदि डेटा श्रृंखला में कोई अवलोकन गायब है तो इसकी गणना नहीं की जा सकती है।

एएम, जीएम और एच.एम के बीच संबंध :

- दिए गए दो प्रेक्षणों के लिए, ए. एम. $\geq$ जी. एम. $\geq$ एच. एम.
- जी. एम. $= \sqrt{\text{ए. एम.} * \text{एच. एम.}}$
- $\text{ए. एम.} = \frac{\text{जी.एम.}^2}{\text{एच.एम.}}$
- $\text{एच. एम.} = \frac{\text{जी.एम.}^2}{\text{ए.एम.}}$

2.2. माध्यिका:

जब सभी प्रेक्षणों को आरोही/अवरोही क्रम में व्यवस्थित किया जाता है तो माध्य मध्य स्थिति में स्थित मान होता है। माध्यिका एक आदेशित डेटा श्रृंखला का केंद्रीय मान है। यह डेटा सेट को ठीक दो भागों में विभाजित करता है। 50 प्रतिशत प्रेक्षण माध्यिका से नीचे हैं और 50% माध्यिका से ऊपर हैं। माध्यिका को 'स्थितीय औसत' के रूप में भी जाना जाता है। मेडियन 50 वाँ प्रतिशतक, 10 वाँ दशमांश और दूसरा चतुर्थक है। माध्यिका तोरण वक्र से कम और अधिक का प्रतिच्छेद बिंदु भी है।

गैर-आवृत्ति डेटा के लिए माध्यिका:

चरण 1 डेटा को सबसे छोटे से सबसे बड़े तक ऑर्डर करें।
चरण 2 यदि प्रेक्षणों की संख्या विषम है, तो $(n + 1)/2^{th}$ वाँ प्रेक्षण (क्रमबद्ध समुच्चय में) माध्यिका है। जब प्रेक्षणों की कुल संख्या सम होती है, तो माध्यिका $n/2th$ और $(n/2 + 1)^{th}$ प्रेक्षण के माध्य से दी जाती है।

समूह आवृत्ति डेटा के लिए माध्यिका:

चरण 1 डेटा के लिए संचयी आवृत्तियों को प्राप्त करें।
चरण 2 उस वर्ग को चिह्नित करें जिसकी संचयी बारंबारता N/2 से अधिक है। वह वर्ग मध्य वर्ग है।
चरण 3 फिर माध्यिका का मूल्यांकन एक प्रक्षेप सूत्र द्वारा किया जाता है

$$Median = l + \frac{h}{f}(\frac{N}{2} - C)$$

जहाँ, l = माध्यिका वर्ग की निचली सीमा
N = अवलोकनों की संख्या
C = माध्यिका वर्ग तक जाने वाले वर्ग की संचयी बारंबारता
f = माध्यिका वर्ग की बारंबारता
h= माध्यिका वर्ग का परिमाण
नोट: आलेखीय रूप से, हम माध्यिका को हिस्टोग्राम द्वारा ज्ञात कर सकते हैं।

माध्यिका का उपयोग:

- गुणात्मक डेटा को परिमाण के आरोही या अवरोही क्रम में व्यवस्थित किया जा सकता है।
- औसत बुद्धि, ईमानदारी आदि का पता लगाएं।

माध्यिका के गुण:

- इसे कड़ाई से परिभाषित किया गया है।
- यह चरम मूल्यों से प्रभावित नहीं है।
- इसे ग्राफिक रूप से स्थित किया जा सकता है।
- इसकी गणना ओपन एंड क्लास बारंबारता वितरण के लिए की जा सकती है।
- इसकी गणना एक क्रमिक पैमाने के आधार पर डेटा के लिए की जा सकती है।

माध्यिका के दोष:

- यह सभी अवलोकनों पर आधारित नहीं है।
- गणना माध्य से अधिक जटिल है।
- यह आगे गणितीय उपचार करने में सक्षम नहीं है।
- माध्य की तुलना में, यह नमूने के उतार-चढ़ाव से बहुत अधिक प्रभावित होता है।

बहुलक:

बहुलक को उस मान के रूप में परिभाषित किया जाता है जो डेटा में सबसे अधिक बार होता है। यदि डेटा सेट में प्रत्येक अवलोकन केवल एक बार होता है, तो इसका बहुलक नहीं होता है। जब डेटा सेट में दो या दो से अधिक मान उच्चतम आवृत्ति के बराबर होते हैं, तो डेटासेट में दो या अधिक मोड मौजूद होते हैं।
अनग्रुप फ़्रीक्वेंसी डेटा के लिए मोड: वह ऑब्जर्वेशन है जिसकी फ़्रीक्वेंसी डेटा सेट में सबसे अधिक होती है।
समूह (समान चौड़ाई) आवृत्ति डेटा के लिए मोड:
चरण 1 मोडल वर्ग की पहचान करें। मोडल क्लास सबसे बड़ी बारंबारता वाला वर्ग है।
चरण 2 प्रक्षेपित सूत्र का उपयोग करके बहुलक ज्ञात कीजिए।

$$mode = l + \frac{h(f_0 - f_{-1})}{(f_0 - f_{-1}) - (f_1 - f_0)}$$

जहाँ, l = बहुलक वर्ग की निचली सीमा
f_0= बहुलक वर्ग की बारंबारता
f_{-1}= पूर्ववर्ती मोडल वर्ग की बारंबारता
f_1= बाद के मोडल वर्ग की बारंबारता
h== बहुलक वर्ग का परिमाण

नोट: आलेखीय रूप से, हम हिस्टोग्राम द्वारा बहुलक ज्ञात कर सकते हैं।

मोड का उपयोग:

- विभिन्न प्रकार के उत्पादों के लिए आदर्श उपभोक्ता वरीयताएँ खोजना।
- जूते या शर्ट के औसत आकार के लिए सबसे अच्छा उपाय।

मोड के गुण:

- यह चरम मूल्यों से प्रभावित नहीं है।
- इसे ग्राफिक रूप से स्थित किया जा सकता है।
- इसकी गणना ओपन एंड क्लास बारंबारता वितरण के लिए की जा सकती है।
- इसकी गणना नाममात्र के पैमाने के आधार पर डेटा के लिए की जा सकती है।

मोड के दोष:

- यह खराब परिभाषित है।
- यह सभी अवलोकनों पर आधारित नहीं है।
- यह आगे गणितीय उपचार करने में सक्षम नहीं है।
- माध्य की तुलना में, यह नमूने के उतार-चढ़ाव से बहुत अधिक प्रभावित होता है।

चतुर्थक: चतुर्थक वे तीन बिंदु हैं जो संपूर्ण डेटा को चार बराबर भागों में विभाजित करते हैं।

$$Q_i = l + \frac{h}{f}(\frac{iN}{4} - C)$$

दशमांश: दशमांश नौ बिंदु हैं जो पूरे डेटा को दस बराबर भागों में विभाजित करते हैं।

$$D_i = l + \frac{h}{f}(\frac{iN}{10} - C)$$

पर्सेंटाइल: पर्सेंटाइल निन्यानवे बिंदु हैं जो पूरे डेटा को सैकड़ों बराबर भागों में विभाजित करते हैं।

$$P_i = l + \frac{h}{f}(\frac{iN}{100} - C)$$

नोट: माध्यिका=दूसरा चतुर्थांश=5वाँ दशमांश=50वाँ शतमक

माध्य माध्यिका और बहुलक के बीच अनुभवजन्य सूत्र: यदि डेटा सेट प्रकृति में असममित हैं, तो

माध्य- बहुलक = 3 (माध्य- माध्यिका)

केन्द्रीय प्रवृति के माप के सबसे अच्छे तरीके:
प्रो. यूल, के अनुसार माध्य केंद्रीय प्रवृत्ति का सर्वोत्तम माप है। लेकिन कुछ स्थितियां ऐसी भी हैं जहां केंद्रीय प्रवृत्ति के अन्य उपायों को प्राथमिकता दी जाती है।

पैमाना	उपयोग माप	पैमाना सर्वोत्तम उपाय
अंतराल	माध्य, माध्यिका, बहुलक	सममित डेटा: माध्य असममित डेटा: माध्यिका
अनुपात	माध्य, माध्यिका,	सममित डेटा: माध्य असममित डेटा: माध्यिका

साधारण	माध्यिका,	माध्यिका
सांकेतिक	बहुलक	बहुलक

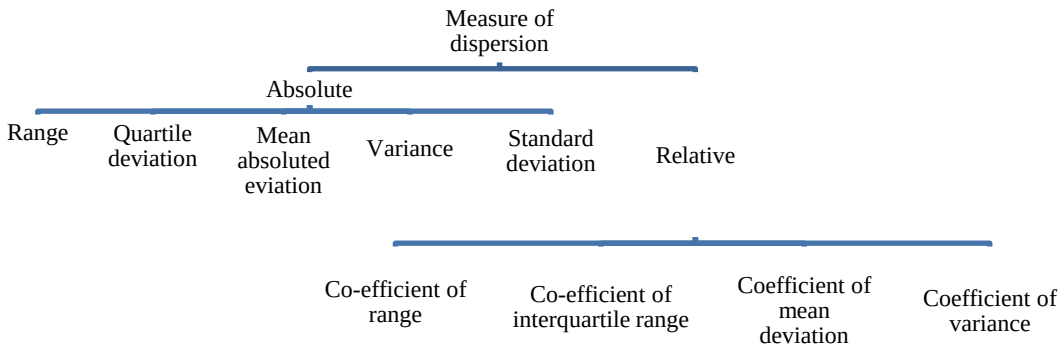
फैलाव का उपाय

केंद्रीय प्रवृत्ति की माप जैसे माध्य, माध्यिका और बहुलक केवल डेटा के केंद्र का पता लगाते हैं। यह डेटा के प्रसार के बारे में कुछ भी अनुमान नहीं लगाता है। दो डेटा सेट का माध्य समान हो सकता है लेकिन वे पूरी तरह से भिन्न हो सकते हैं।

डेटा 1	38	42	41	44	45
डेटा 2	50	53	41	35	31

उपरोक्त उदाहरण में, दो डेटासेट का माध्य समान है। अतः केन्द्रीय प्रवृत्ति के माप आँकड़ों का वर्णन करने के लिए पर्याप्त नहीं हैं। इस प्रकार आँकड़ों का वर्णन करने के लिए प्रेक्षणों के प्रकीर्णन के माप को जानना आवश्यक है। फैलाव को उनके केंद्रीय मूल्यों से विचलन या प्रेक्षणों के बिखराव के रूप में परिभाषित किया गया है।

परिक्षेपण के माप के विभिन्न तरीके :



रेंज (आर):

रेंज फैलाव का सबसे सरल उपाय है। इसे चर के उच्चतम मान और निम्नतम मान के बीच के अंतर के रूप में परिभाषित किया गया है। यह फैलाव का एक कच्चा उपाय है।

रेंज = उच्चतम मूल्य (एच) - सबसे कम मूल्य (एल)

रेंज के गुण:

- इसे समझना और गणना करना आसान है।
- यह डेटा की आवृत्ति से प्रभावित नहीं होता है।

रेंज के दोष:

- यह सभी अवलोकनों पर निर्भर नहीं करता है।
- यह चरम वस्तुओं से बहुत अधिक प्रभावित होता है।
- इसकी गणना ओपन-एंड क्लास अंतराल से नहीं की जा सकती है।
- यह आगे के गणितीय उपचार के लिए उपयुक्त नहीं है।
- यह फैलाव का सबसे अविश्वसनीय उपाय है।

चतुर्थक विचलन (Q.D.):
इंटरक्वेर्टाइल रेंज पहले और तीसरे क्वार्टाइल के बीच का अंतर है। इसलिए इंटरक्वेर्टाइल रेंज 50% टिप्पणियों के मध्य का वर्णन करती है।

इंटर क्वार्टाइल रेंज = Q3- Q1

जहाँ पे,
Q^3= डेटा का पहला चतुर्थक
Q^1= डेटा का तीसरा चतुर्थक
चतुर्थक विचलन (Q.D.) अंतःचतुर्थक श्रेणी का आधा है।

$$\text{चतुर्थक विचलन } (\boldsymbol{Q.D.}) = \frac{\boldsymbol{Q3 - Q1}}{\boldsymbol{2}}$$

चतुर्थक विचलन के गुण:

- इसे समझना और गणना करना आसान है।
- यह चरम मूल्यों से प्रभावित नहीं है।
- इसकी गणना ओपन एंड फ़्रीक्वेंसी डेटा के लिए की जा सकती है।

चतुर्थक विचलन के दोष:

- यह सभी अवलोकनों पर निर्भर नहीं करता है।
- यह आगे के गणितीय उपचार के लिए उपयुक्त नहीं है।
- यह नमूने के उतार-चढ़ाव से बहुत अधिक प्रभावित होता है।

मीन निरपेक्ष विचलन (एमएडी):
केंद्रीय मान (माध्य बेहतर है) से प्रत्येक मान के निरपेक्ष विचलन की गणना की जाती है और इन विचलनों के अंकगणितीय माध्य को माध्य निरपेक्ष विचलन कहा जाता है।
गैर-आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \ldots, X_n$ डेटा सेट के n अवलोकन हैं। A के बारे में माध्य निरपेक्ष विचलन (MAD) द्वारा दिया गया है

$$MAD_A = \frac{\sum_{i=1}^{n} |X_i - A|}{n}$$

माध्य के बारे में माध्य निरपेक्ष विचलन (MAD) द्वारा दिया जाता है

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} |X_i - \bar{X}|}{n}$$

आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \ldots, X_n$ संगत आवृत्तियों के साथ अवलोकन हैं $f_1, f_2, \ldots, f_n$ तथा $\sum_{i=1}^{n} f_i = N$. A के बारे में माध्य निरपेक्ष विचलन (MAD) द्वारा दिया गया है

$$MAD_A = \frac{\sum_{i=1}^{n} f_i |X_i - A|}{N}$$

माध्य के बारे में माध्य निरपेक्ष विचलन (MAD) द्वारा दिया जाता है

$$MAD_{\bar{X}} = \frac{\sum_{i=1}^{n} f_i |X_i - \bar{X}|}{N}$$

माध्य के सापेक्ष माध्य निरपेक्ष विचलन के गुण:

- इसे समझना और गणना करना आसान है।
- यह सभी अवलोकनों पर आधारित है।

माध्य के सापेक्ष माध्य निरपेक्ष विचलन के अवगुण:

- यह आगे के गणितीय उपचार के लिए उपयुक्त नहीं है।
- यह विचलन के संकेत को ध्यान में नहीं रखता है।
- यह चरम मूल्यों से प्रभावित है।

मानक विचलन (एस.डी.):

यह फैलाव का सबसे अच्छा उपाय और सबसे अधिक इस्तेमाल किया जाने वाला उपाय है। इसे उनके अंकगणितीय माध्य से दिए गए अवलोकन के विचलन के वर्ग के अंकगणितीय माध्य के धनात्मक वर्गमूल के रूप में परिभाषित किया गया है। यह सभी प्रेक्षणों के परिमाण को ध्यान में रखता है और संभावित फैलाव का न्यूनतम मान देता है। इसे माध्य के बारे में मूल माध्य वर्ग विचलन के रूप में भी जाना जाता है।

गैर-आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \ldots, X_n$ डेटा सेट के n अवलोकन हैं। मानक विचलन A द्वारा दिया गया है

$$\text{एस. डी.} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n}}$$

आवृत्ति डेटा के लिए: मान लें कि $X_1, X_2, \ldots, X_n$ संगत आवृत्तियों के साथ अवलोकन हैं $f_1, f_2, \ldots, f_n$ तथा $\sum_{i=1}^{n} f_i = N$. मानक विचलन द्वारा दिया जाता है

$$\text{एस. डी.} = \sqrt{\frac{\sum_{i=1}^{n} f_i(X_i - \bar{X})^2}{N}}$$

मानक विचलन के गुण:

- यह उत्पत्ति के परिवर्तन से स्वतंत्र है लेकिन पैमाने के परिवर्तन पर निर्भर है
  let $U = a + hX$, then sd(U)=|h|*sd(x)
- यदि सभी प्रेक्षण समान हैं तो मानक विचलन शून्य है।
- यह कभी भी चतुर्थक विचलन और माध्य निरपेक्ष विचलन से कम नहीं होता है।

मानक विचलन के गुण:

- यह सभी अवलोकनों पर आधारित है।
- यह चरम मूल्यों से कम प्रभावित होता है।
- यह आगे के गणितीय उपचार के लिए उपयुक्त है।

मानक विचलन के दोष:

- यह आगे के गणितीय उपचार के लिए उपयुक्त है।
- यह विचलन के संकेत को ध्यान में नहीं रखता है।
- यह चरम मूल्यों से प्रभावित है।
- ओपन-एंड क्लास डेटा के लिए इसकी गणना नहीं की जा सकती है।

वरिएंस

इसे मानक विचलन के वर्ग के रूप में परिभाषित किया गया है। प्रसरण की इकाई वास्तविक प्रेक्षणों का वर्ग है, जबकि मानक विचलन की इकाई वास्तविक प्रेक्षणों के समान है।

Relations between R, Q.D., M.D. and S.D.

$$\mathbf{9QD = \frac{15}{2}MD = 6SD = R}$$

भिन्नता का गुणांक (C.V.):

डेटा सेट के लिए भिन्नता के गुणांक को मानक विचलन के अनुपात के रूप में परिभाषित किया गया है और प्रतिशत में व्यक्त किया गया है।

$$\boldsymbol{CV = \frac{SD}{mean} * 100\%}$$

C.V परिक्षेपण का सापेक्ष माप है। यह फैलाव के सभी सापेक्ष मापों में सबसे अच्छा उपाय है। C.V का उपयोग दो या दो से अधिक डेटा श्रृंखलाओं के बीच परिवर्तनशीलता या स्थिरता की तुलना करने के लिए किया जाता है। यदि सी.वी. अधिक है यह दर्शाता है कि समूह अधिक परिवर्तनशील, कम स्थिर, कम एकसमान और कम

सुसंगत है। यदि सी.वी. कम है, यह इंगित करता है कि समूह कम परिवर्तनशील या अधिक स्थिर या अधिक एकसमान और अधिक सुसंगत है।

उदाहरण: एकदिवसीय क्रिकेट में कोहली और स्मिथ के स्कोर के आंकड़ों पर विचार करें। कोहली के लिए माध्य और मानक विचलन क्रमशः 55 और 5 हैं। स्मिथ के लिए माध्य और मानक विचलन क्रमशः 50 और 10 हैं। सी.वी. खोजें दोनों डेटा के लिए मूल्य और उनकी तुलना करें।

समाधान:

कोहली के लिए, CV=5/55*100=9%

स्मिथ के लिए, CV=10/50*100=20%

स्मिथ कोहली की तुलना में स्कोर में अधिक भिन्नता के अधीन है। इसलिए कोहली स्मिथ से ज्यादा सुसंगत हैं।

3.6. Coefficient of range $= \frac{H-L}{H+L} * 100\%$

3.7. Coefficient of inter quartile range $= \frac{Q3-Q1}{Q3+Q1} * 100\%$

3.8. Coefficient of mean deviation $= \frac{MAD}{averave\ from\ which\ it\ is\ calculated} * 100\%$

संख्यात्मक उदाहरण: सांख्यिकी परीक्षा में 10 छात्रों के अंक इस प्रकार हैं:

10,12,15,12,16,20,13,17,15,15

माध्य, माध्यिका, बहुलक, परास और मानक विचलन ज्ञात कीजिए।

समाधान:

X_i	f_i	$f_i X_i$	$f_i (X_i - \overline{X})$	$(X_i - \overline{X})^2$	$f_i (X_i - \overline{X})^2$
10	1	10	-4.5	20.25	20.25
12	2	24	-5	6.25	12.5
13	1	13	-1.5	2.25	2.25
15	3	45	1.5	0.25	0.75
16	1	16	1.5	2.25	2.25
17	1	17	2.5	6.25	6.25
20	1	20	5.5	30.25	30.25
Total	10	145		67.75	74.5

$$\text{माध्य} = \frac{145}{10} = 14.5$$

$$\text{माध्यिका} = 15$$

$$\text{मोड} = 15$$

रेंज=20-10=10

$$\text{एस. डी.} = \frac{74.5}{10} = 7.45$$

4. स्क्यूनिस और कुर्टोसिस:

हमने केंद्रीय प्रवृत्ति के उपायों और फैलाव के माप पर चर्चा की है जो डेटा सेट के स्थान और स्केल पैरामीटर का वर्णन करते हैं। वे डेटा संरचना के आकार के बारे में कोई विचार नहीं देते हैं। स्क्यूनिस और कर्टोसिस का माप डेटा सेट के आकार को दर्शाता है। स्क्यूनिस का माप समरूपता की कमी की दिशा और परिमाण देता है और कुर्टोसिस का माप वक्र की समतलता का विचार देता है।

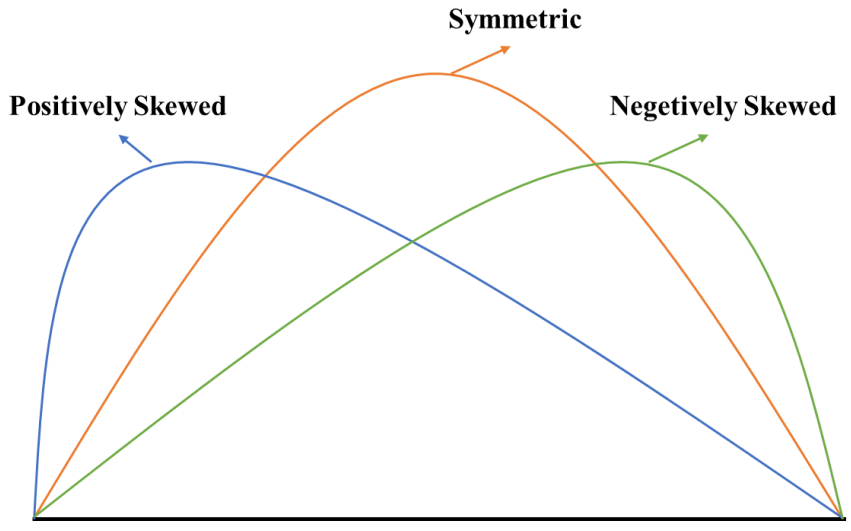
4.1 स्क्यूनिस

स्क्यूनिस डेटा की विषमता की डिग्री को मापता है। स्क्यूनिस समरूपता की कमी को दर्शाता है।
स्क्यूनिस मुख्य रूप से तीन प्रकार का होता है: सकारात्मक स्क्यूनिस, नकारात्मक स्क्यूनिस और सममित डेटा।

सकारात्मक स्क्यूनिस:
एक डेटा को सकारात्मक स्क्यूनिस कहा जाता है यदि लंबी पूंछ शिखर के दाईं ओर हो। माध्य शिखर मान के दाईं ओर है। यहाँ माध्य > माध्यिका > बहुलक है।

नकारात्मक स्क्यूनिस:
एक डेटा को नकारात्मक स्क्यूनिस कहा जाता है यदि लंबी पूंछ शिखर के बाईं ओर है। माध्य शिखर मान के बाईं ओर है। यहाँ माध्य <माध्यिका <मोड।

सममित
सममित वितरण में शून्य विषमता होती है क्योंकि केंद्रीय प्रवृत्ति के सभी माप मध्य में होते हैं। जब डेटा सममित रूप से वितरित किया जाता है, तो बाईं ओर और दाईं ओर समान संख्या में अवलोकन होते हैं। यहाँ माध्य = माध्यिका = बहुलक।



चित्र 1. स्क्यूनिस

The measure of Skewness:

$$\text{Pearson's first coefficient} = \frac{\text{Mean} - \text{Mode}}{\text{Standard Deviation}}$$

$$\text{Pearson's second coefficient} = \frac{3\,(\text{Mean} - \text{Median})}{\text{Standard Deviation}}$$

स्क्यूनिस का माप:
व्याख्या:
1. यदि Sk = 0 है, तो बारंबारता बंटन सामान्य और सममित होता है।
2. यदि Sk > 0, तो बारंबारता बंटन धनात्मक रूप से विषम होता है।
3. यदि Sk <0 है, तो बारंबारता बंटन ऋणात्मक रूप से विषम होता है।

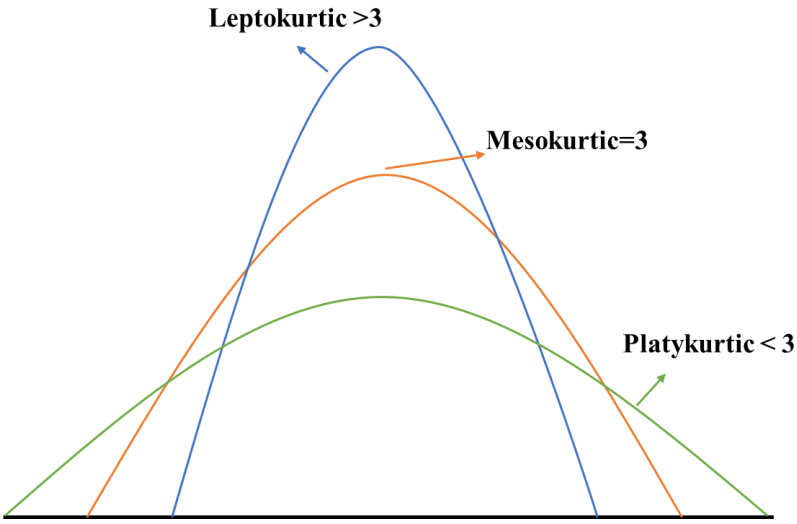
कर्टोसिस

कर्टोसिस एक माप है कि क्या डेटा सामान्य वितरण के सापेक्ष भारी-पुच्छ या हल्के-पुच्छ हैं। यही है, उच्च कर्टोसिस वाले डेटा सेट में भारी पूंछ या आउटलेयर होते हैं। कम कर्टोसिस वाले डेटा सेट में हल्की पूंछ या आउटलेर्स की कमी होती है। एक समान वितरण चरम मामला होगा।
कर्टोसिस के प्रकार: लेप्टोकोर्टिक या हेवी-टेल्ड डिस्ट्रीब्यूशन, मेसोकोर्टिक, प्लेटीकुर्टिक या शॉर्ट-टेल्ड डिस्ट्रीब्यूशन

लेप्टोकुर्टिक
लेप्टोकोर्टिक इंगित करता है कि वितरण चरम पर है और इसमें मोटी पूंछ है।

प्लेटीकुरटिक
प्लेटीकुर्टिक में निचली पूंछ होती है और केंद्र की पूंछ के चारों ओर फैली हुई होती है, इसका मतलब है कि अधिकांश डेटा बिंदु माध्य के साथ उच्च निकटता में मौजूद हैं। सामान्य वितरण के साथ तुलना करने पर एक प्लेटीकुर्टिक वितरण समतल (कम चोटी वाला) होता है।

मेसोकुर्टिक
मेसोकर्टिक सामान्य वितरण के समान है। मेसोकोर्टिक में, वितरण चौड़ाई में मध्यम होते हैं, और वक्र मध्यम चोटी की ऊंचाई होते हैं।



Measurement of Kurtosis $(\beta_2) = \frac{1}{N-1}\frac{\sum(y_i-\bar{y})^4}{s^4}$

$\gamma_2 = \beta_2 - 3$

5. डेटा की प्रस्तुति

डेटा प्रस्तुत करने के तीन व्यापक तरीके हैं। ये हैं टेक्स्ट प्रेजेंटेशन, टेबुलर प्रेजेंटेशन और ग्राफिक या डायग्रामेटिक प्रेजेंटेशन। हमने आँकड़ों की केवल कुछ महत्वपूर्ण आरेखीय प्रस्तुतियों पर चर्चा की।

गैर आयामी आरेख	चित्रलेख
द्विविमीय आरेख	दंड आरेख, पाई आरेख, हिस्टोग्राम, बॉक्स प्लॉट
त्रिविमीय आरेख	घन, सिलिंडर आरेख

5.1 बार आरेख

5.1.1 सरल बार आरेख

यदि वर्गीकरण विशेषताओं पर आधारित है और यदि विशेषताओं की तुलना किसी एकल वर्ण से की जानी है तो हम एक साधारण दंड आरेख का उपयोग करते हैं। सरल दंड आरेखों में समान चौड़ाई के ऊर्ध्वाधर दंड होते हैं। इन पट्टियों की ऊँचाई गुण के आयतन या परिमाण के समानुपाती होती है। सभी बार एक ही आधार रेखा पर खड़े होते हैं। सलाखों को एक दूसरे से समान अंतराल से अलग किया जाता है। सलाखों को रंगीन या चिह्नित किया जा सकता है।

5.1.2 एकाधिक दंड आरेख

यदि डेटा को विशेषताओं द्वारा वर्गीकृत किया जाता है और यदि प्रत्येक विशेषता के भीतर दो या दो से अधिक वर्णों या समूहों की तुलना की जानी है तो हम कई बार आरेखों का उपयोग करते हैं। यदि प्रत्येक विशेषता के भीतर केवल दो वर्णों की तुलना की जाती है, तो परिणामी दंड आरेख का उपयोग किया जाता है जिसे दोहरा दंड आरेख के रूप में जाना जाता है। बहु दंड आरेख एक साधारण दंड आरेख का ही विस्तार है। प्रत्येक विशेषता के लिए, अलग-अलग वर्णों या समूहों का प्रतिनिधित्व करने वाले दो या दो से अधिक बार एक साथ रखे जाने हैं। एक विशेषता के भीतर प्रत्येक बार को अलग करने के लिए अलग-अलग तरीके से चिह्नित या रंगीन किया जाएगा। प्रत्येक विशेषता के तहत एक ही प्रकार की मार्किंग या रंगाई की जानी चाहिए। चिह्नों या रंगों की व्याख्या करते हुए एक फुटनोट दिया जाना चाहिए।

5.1.3 घटक दंड आरेख

इसे उपविभाजित दंड आरेख भी कहते हैं। प्रत्येक घटक के लिए सलाखों को एक साथ रखने के बजाय, हम इसे एक के ऊपर एक रख सकते हैं। इसका परिणाम एक घटक बार आरेख होगा।

5.2. हिस्टोग्राम

हिस्टोग्राम निरंतर वर्ग बारंबारता वितरण के लिए उपयुक्त है। हम x-अक्ष के साथ वर्ग अंतरालों और y-अक्ष के अनुदिश आवृत्तियों (असमान आवृत्ति डेटा के लिए बारंबारता घनत्व) को चिह्नित करते हैं।

- समान वर्ग अंतराल के लिए, आयतों की ऊँचाई बारंबारता के समानुपाती होगी, जबकि असमान वर्ग अंतरालों के लिए ऊँचाई बारंबारता घनत्व के बराबर (या आनुपातिक) होगी।
- एक बारंबारता बहुभुज हिस्टोग्राम में आयतों के शीर्ष के मध्य बिंदुओं को जोड़कर प्राप्त किया गया एक रेखा ग्राफ है।

तालिका 1. दंड आरेख और हिस्टोग्राम के बीच अंतर

अभिलक्षण	बार आरेख	हिस्टोग्राम
बारंबारता को मापा जाता है	बार की ऊंचाई	बार का क्षेत्रफल
बार के बीच का गैप	हां	नहीं
बार की चौड़ाई	बराबर	बराबर नहीं हो सकती
डेटा प्रकार	केवल असतत और सतत	निरंतर

5.3. पाई आरेख

जब हम एक कारक के विभिन्न घटकों के सापेक्ष महत्व में रुचि रखते हैं, तो हम पाई आरेखों का उपयोग करते हैं। पाई आरेख के लिए, एक वृत्त का उपयोग किया जाता है और इसके द्वारा घेरे गए क्षेत्र को 100 के रूप में लिया जाता है। फिर इसे केंद्र में कोण बनाकर कई क्षेत्रों में विभाजित किया जाता है, प्रत्येक क्षेत्र का क्षेत्रफल संबंधित प्रतिशत का प्रतिनिधित्व करता है।

5.4. बॉक्स प्लाट

न्यूनतम, अधिकतम और चतुर्थक (Q1, माध्यिका, Q3) एक साथ एक अच्छे कॉम्पैक्ट तरीके से चर के केंद्र और भिन्नता के बारे में जानकारी प्रदान करते हैं। बढ़ते क्रम में लिखा गया है, वे चर के पांच-संख्या सारांश कहलाते हैं। एक बॉक्सप्लॉट पांच-नंबर सारांश पर आधारित होता है और इसका उपयोग केंद्र के ग्राफिकल डिस्प्ले और डेटा सेट में चर के देखे गए मानों की भिन्नता प्रदान करने के लिए किया जा सकता है। यह आपको आपके आउटलेर्स और उनके मूल्यों के बारे में बता सकता है। यह आपको यह भी बता सकता है कि आपका डेटा सममित है या नहीं, आपका डेटा कितनी मजबूती से समूहीकृत है, और आपका डेटा विषम है या नहीं।

N.B: एक्सेल मैनुअल के साथ हमारे बेसिक स्टैटिस्टिक्स में ग्राफिकल प्रेजेंटेशन के उदाहरण दिए गए हैं।

6. माध्य और मानक विचलन का मजबूत अनुमान

माध्य और मानक विचलन केवल तभी सही अनुमान प्रदान करता है जब चर सामान्य रूप से वितरित किया जाता है और बाहरी कारकों के बिना। यदि चर विषम है और/या आउटलेयर है, तो माध्य और मानक विचलन अत्यधिक अवलोकनों से अत्यधिक प्रभावित होंगे और डेटा के दोषपूर्ण आंकड़े प्रदान करेंगे। माध्य और मानक विचलन के कई विकल्प हैं। माध्य के विकल्प में सुप्रसिद्ध माध्यिका और छंटे हुए माध्य, विंसोराइज़्ड माध्य और M-आकलनकर्ता शामिल हैं और मानक विचलन के लिए विकल्पों में अंतर-चतुर्थक श्रेणी (IQR) और माध्य निरपेक्ष विचलन (MAD), छंटे हुए मानक विचलन शामिल हैं। Winsorized मानक विचलन, और एम-आकलनकर्ता। माध्यिका, IQR, MAD की चर्चा पिछले सेकंड में की जा चुकी है

6.1. छंटे हुए माध्य और मानक विचलन

एक छंटे हुए माध्य और मानक विचलन एक "नियमित" माध्य के समान हैं, लेकिन यह दोनों तरफ से किसी भी आउटलेयर को ट्रिम कर देता है। 20% छंटे हुए माध्य को प्राप्त करने के लिए, 20% न्यूनतम और 20% उच्चतम मूल्यों को हटा दिया जाता है और शेष टिप्पणियों पर माध्य की गणना की जाती है। हमारे उदाहरण में, ये मान होंगे: 4, 4, 5, 5, 6, 6, और 20% छंटनी का मतलब 5 के बराबर होगा।

6.2. Winsorized माध्य और मानक विचलन

Winsorized तकनीक ट्रिम की गई तकनीक के समान है लेकिन सबसे कम (resp.highest) मानों को हटाया नहीं जाता है बल्कि सबसे कम (resp.highest) untrimmed स्कोर द्वारा प्रतिस्थापित किया जाता है। हमारे उदाहरण में, वेरिएबल्स का मान, जिसे विन्सोराइज़्ड स्कोर भी कहा जाता है, तब होगा: 4, 4, 4, 4, 5, 5, 6, 6, 6, 6, और 20% विन्सोराइज़्ड माध्य 5 के बराबर होगा। .

6.3. एम अनुमानक

छंटे हुए माध्य सभी या तो अवलोकन लेते हैं या छोड़ते हैं। Winsorized माध्य के लिए, यह कम चरम मानों वाले मानों को प्रतिस्थापित करता है। इसके विपरीत, एम-आकलनकर्ता, प्रत्येक अवलोकन को उसके विशेष गुणों के लिए चयनित फ़ंक्शन के अनुसार भारित करते हैं। भार एक स्थिरांक पर निर्भर करता है जिसे

शोधकर्ता द्वारा चुना जा सकता है। एम-आकलनकर्ता प्रेक्षणों को उत्तरोत्तर कम करके कई प्रेक्षणों को शून्य मान निर्दिष्ट करने की इस समस्या को हल करता है। एम-आकलनकर्ता का एकमात्र पहलू जो वास्तविक शोधकर्ताओं को चिंतित कर सकता है, वह यह है कि किसी को टिप्पणियों के डाउनवेटिंग की डिग्री का चयन करना चाहिए।

एमएस एक्सेल का उपयोग कर सांख्यिकीय तकनीक

मो यासीन

भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012

md.yeasin@icar.gov.in

माइक्रोसॉफ्ट एक्सैलसदैव उपयोग किये जाने वाला एक सॉफ्टवेयर अनुप्रयोग है, विश्वभर में लाखों लोगमाइक्रोसॉफ्टएक्सैलका प्रयोग करते हैं। प्रयोक्ता एक्सैलमें हर प्रकार के आंकड़ों की प्रवृष्टि कर सकते हैं तथा वित्तीय, गणतीय एवं सांख्यिकीय गणनायें कर सकते हैं ।

माइक्रोसॉफ्ट एक्सैलआंकड़ों के विश्लेषण के लिए एनालिसिस टूल पैक प्रदान करता है जिसकी सहायता से जटिल सांख्यिकीय अथवा अभियांत्रिकीय विश्लेषण के विभिन्न चरणों कोबचाया (save) जा सकता है तथा प्रत्येक विश्लेषण के लिए आंकड़े एवं प्राचल प्रदान करता है । इस औजार की सहायता से उपयुक्त सांख्यिकीय अथवा अभियांत्रकीय सूक्ष्म गणनायें की जा सकती हैं तथा प्राप्त परिणामों को तालिका में दर्शाया जाता सकता है । कुछ औजार परिणामों की तालिका के अतिरिक्त चार्ट का भी सृजन करते हैं ।

1.स्वतन्त्र प्रतिदर्श t-जांच

दो जनसंख्या औसतों का सांख्यिकीय परीक्षण द्विप्रतिदर्श t-जांच की सहायता से यह जानने के लिये परीक्षण किया जाता है कि क्या दोनों प्रतिदर्श भिन्न हैं, साथ ही जब दोनों सामान्य वितरणों के प्रसरण अज्ञात हों तथा परीक्षण प्रतिदर्श का आकार छोटा हो तो ऐसी परिस्थिति में भी द्विप्रतिदर्श t-जांच का उपयोग किया जाता है ।

निम्नलिखित उदाहरण की सहायता से एक्सेल मे t-जांच की विधि को समझा जा सकता है। t-जांच को शून्य प्राकल्पनाकी जांच के लिए प्रयोग किया जाता है जिसका अभिप्राय है कि दोनों जनसंख्याओं के औसत बराबर है । निम्नलिखत उदाहरण मे 6 महिला विद्यार्थियों तथा 5 पुरुष विद्यार्थियों के अध्ययन के घंटों को लिया गया है ।

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

H15 f_x

	A	B	C
1	Female	Male	
2	26	23	
3	25	30	
4	43	18	
5	34	25	
6	18	28	
7	52		
8			
9			

t-जांच के लिए निम्नलिखित चरणों का पालन करें ।

1.सर्वप्रथम यह जानने के लिए कि क्या दोनों जनसंख्याओं के प्रसरण समान है। F-जांच करें ।

(क) डाटा टैब पर 'डाटा एनालिसिस बटन' को क्लिक करें ।

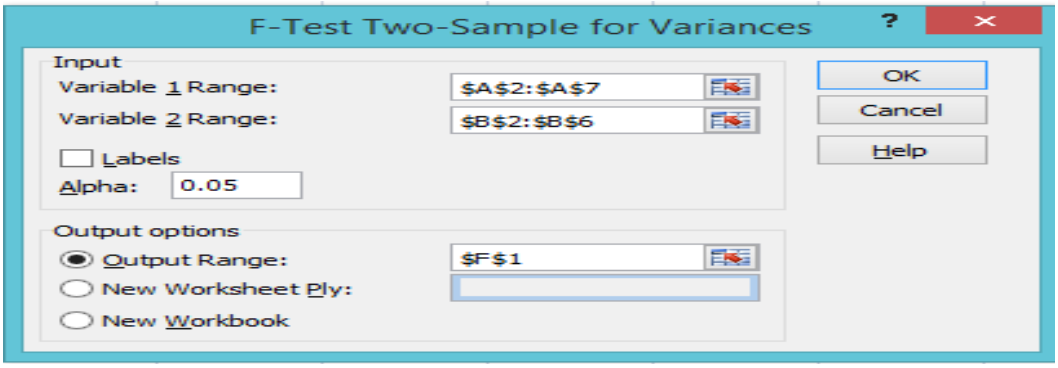
नोट– यदि डाटा एनालिसिस बटन न हो तो 'एनालिसिस टूल पैक' को लोड करें ।

(ख) प्रसरण के लिए द्वि प्रतिदर्श F-जांच को चुनें

(ग) वैरीयेबल–1 रेंज बाक्स को क्लिक करें तथा A_2:A_7 रेंज को चुनें ।

(घ) वैरीयेबल–2 रेंज बाक्स को क्लिक करें तथा B_2:B_6 रेंज को चुनें ।

(च) आउटपुट रेंजबाक्स को क्लिक करें तथा F_1क्लिक करें ।

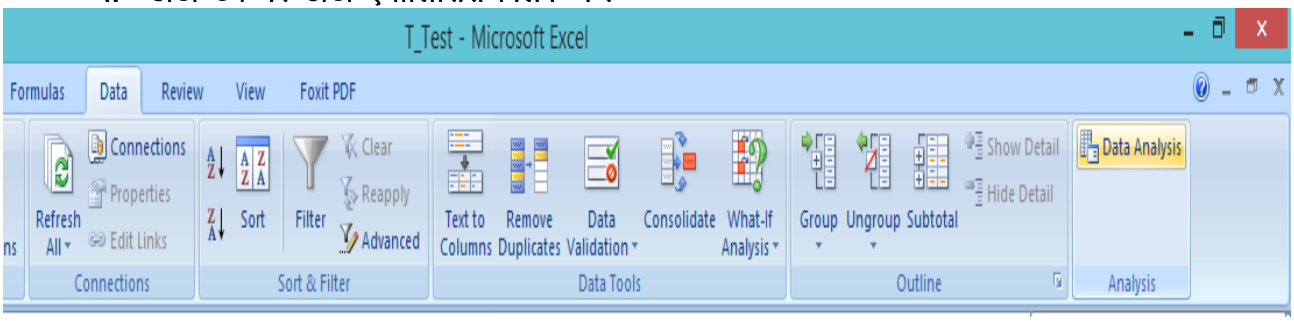


(छ) ओके क्लिक करें ।

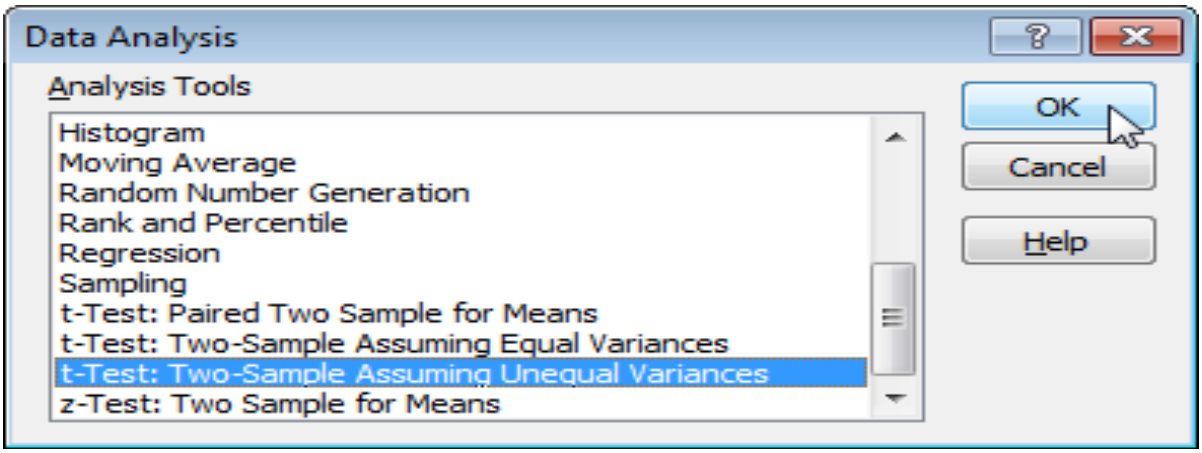
	F	G	H
F-Test Two-Sample for Variances			
		Variable 1	*Variable 2*
Mean		33	24.8
Variance		160	21.7
Observations		6	5
df		5	4
F		7.373272	
P(F<=f) one-tail		0.037888	
F Critical one-tail		6.256057	

F (7.373272) F क्रांतिक (6.256057) से बड़ा है । अतः दोनों जनसंख्याओं के प्रसरण भिन्न हैं। यह एक भिन्न प्रसरणों की अवस्था है

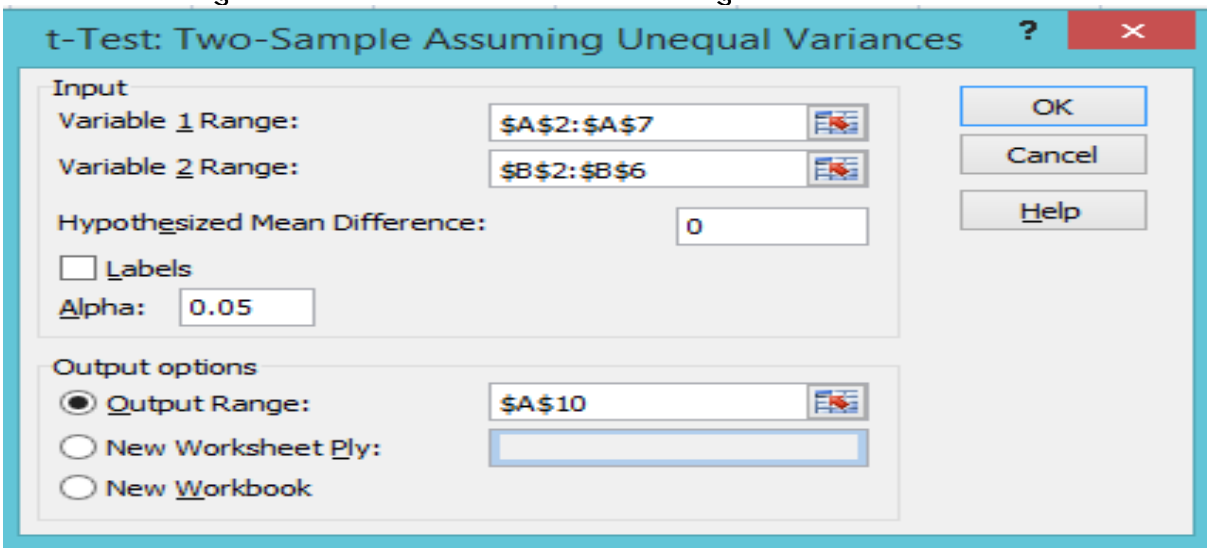
1. डाटा टैब पर डाटा एनालिसिस क्लिक करें



2. t-टेस्ट : टू सैंपल एज्यूमिंग अनईक्वल वैरीयेसेंस को चुनें तथा ओके क्लिक करें ।



2. वैरीयेबल –1 रेंज बाक्स को क्लिक करें तथा रेंजA2:A7 को चुनें ।
3. वैरीयेबल–2 रेंज को क्लिक करें तथा रेंज B2:B6 को चुनें ।
4. हाइपोथाइज्ड मीन डिफरेंस बाक्स को क्लिक करें तथा '0' टाइप करें ($H_0 : \mu_1 - \mu_2 = 0$)।
5. आउटपुट रेंजबाक्स को क्लिक करें तथा E10 को चुनें ।



6- ओके क्लिक करें

परिणामः

t-जांचः असमान प्रसरण मानते हुए प्रतिदर्श

	पुरुष	महिला
औसत	33	24-8
प्रसरण	160	21-7
प्रेक्षण	6	5
प्राकल्पित औसत		
अन्तर	0	
स्वतंत्रता की कोटि	7	
t-सांख्यिकी	1.472605	

एकल पुच्छ	0.09217	
क्रान्तिक एकल पुच्छ	1.894579	
क्रान्तिक द्वि. पुच्छ	0.18434	
t- क्रान्तिक द्वि. पुच्छ	2.364624	

निष्कर्षः हम एक द्वि. पुच्छ जांच (असमानता) करते हैं । यदि t- सांख्यिकी <t क्रान्तिक द्वि. पुच्छ अथवा t- सांख्यिकी >tक्रान्तिक द्वि. पुच्छ हों तो हम शून्य प्राकल्पनाको नकार देते हैं परन्तु यहां ऐसा नहीं है । यहां – 2.365 < 1.473<2.365 है ।इसलिए शून्य प्राकल्पनाको नकारा नहीं जा सकता है । प्रतिदर्श औसतों (3.3-24.8) के बीच के अंतर से स्पष्ट नहीं होता है कि महिला और पुरुष विद्यार्थियों के अध्ययन के घंटों में कोई महत्वपूर्ण अंतर है ।

2.युगल **t**-जांचः.युगल t-जांच प्रायः प्रतिदर्श समूहों के प्राप्ताकों की हस्तक्षेप से पूर्व अथवा पश्चात तुलना के लिए किया जाता है । युगल t-जांच को दो जनसंख्या औसतों की तुलना करने के लिए किया जाता हे । जब हमारे पास दो प्रतिदर्श हों तथा एक प्रतिदर्श के प्रेक्षणों को दूसरे प्रतिदर्श के प्रेक्षणों के साथ जोड़ा बनाया जा सके ।

एक्सैल में युगल **t**-जांच :

दो युगल मानों (जैसे किसी स्थिति से पूर्व अथवा पश्चात) की तुलना के लिए, जब दोनों प्रक्षण एक ही वस्तु अथवा अनुरुप वस्तुओं से लिये गये हों, ऐसी अवस्था में युगल t-जांच का प्रयोग किया जा सकता है । उदाहरण के लिए हमारे पास 8 वस्तुओं के आंकड़ों के दो चर पूर्व और पश्चात हों (आहार से पूर्व एवं पश्चात आर) ।

जांच की प्राकल्पनाइस प्रकार हैः

H_0 : m loss = 0भार में औसत कमी शून्य थी)

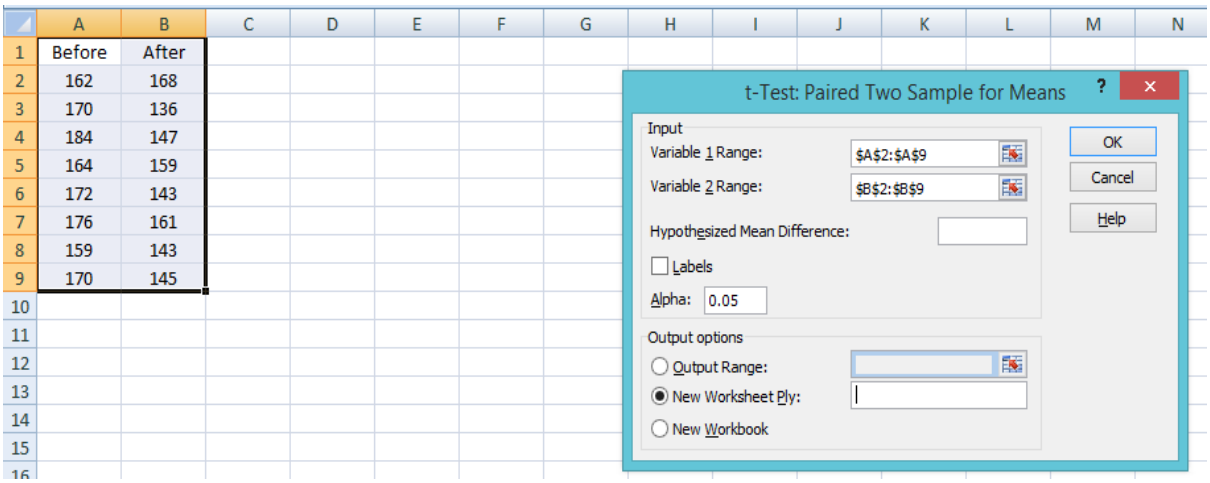
H_a : m loss $\neq$ 0 (भार में औसत कमी शून्य से भिन्न थी)

उदाहरण के लिए भार में कमी के निम्नलिखित आंकड़ों को युगल t-जांच के लिए गया है।

DIET.XLS

पूर्व	पश्चात
162	168
170	136
184	147
164	159
172	143
176	161
159	143
170	145

1. युगल t-जांच के लिए टूल्स डाटा एनालिसिसि /t-टेस्ट: पेर्यड टू सैम्पल फार मीन्सको चुनें ।
2. t-टेस्ट: पेर्यड टू सैम्पल फार मीन्सडायलॉग बाक्स मेंचर –1 की इनपुट रेंज के लिए समूह पूर्वमें भार के 38 मानों को हाइलाइट करें (162 से 170 के मान) । चर–2 की इनपुट रेंज के लिए समूह पश्चातमें भार के 8 मानों को हाइलाइट करें (168 से 145 के मान) । अब अन्य वस्तुओं की उनकी डिफाल्ट अवस्था में छोड़ दें । यहां डायलॉग बाक्स दिखाया गया है । ओके क्लिक करें ।



3. परिणामों को निम्नलिखित आउटपुट तालिका में दशार्या गया है ।

t-जांचः औसतों के लिए युग्मिक द्वि–प्रतिदर्श

	चर–1	चर–2
औसत	169.625	150.25
प्रसरण	65.125	121.9286
प्रेक्षण	8	8
पियरसन सहसंबंध	-0.17675	
परिकाल्पित औसत		
अंतर	0	
स्वतंत्रता की कोटि	7	
t-सांख्यिकी	3.706873	
P(t< = t)एकल पुच्छ	0.003793	
tक्रान्तिक एकल पुच्छ	1.894579	
P(t< = t)द्वि पुच्छ	0.007586	
द्वि पुच्छ	2.364624	

अतः इस t जांच के लिए द्वि पुच्छ P मान है P = 0-008 (0-00758तथा t = 3.71
इस जांच के परिणामों से हम वह प्राप्त नहीं कर सके जो हम वास्तव में चाहते हैं । इसे भली भांति समझने के लिए हमें यह जानना आवश्यक है कि युगल t-जांच वास्तव में दो मानों के बीच के अन्तर की जांच है । अतः एक बेहतर विश्लेषण के लिए पहले पूर्व एवंपश्चात के मानों का अंतर निकालना चाहिए । इसके लिए एक अतिरिक्त स्तम्भ अन्तर का सृजन सूत्र =A2-B2के प्रयोग से किया गया है तथा सूत्र को शेष सभी सेल में कापी किया गया है । औसत अन्तर की भी गणना की गई है ।

पूर्व	पश्चात	अन्तर
162	168	&6
170	136	34

184	147	37
164	159	5
172	143	29
176	161	15
159	143	16
170	145	25

औसत अन्तर = 19.375

यदि हम मूल प्राकल्पना को देखें तो इसमें औसत का मान शून्य से भिन्न है ।

इस प्रकार t-जांच वास्तव में यह जांच कर रहीं है कि क्या 19.38,इसकी उपयोगिता के दावे के लिए प्रर्याप्त मात्रा में शून्य से भिन्न14 चज हे । अतः हम औसत अंतर (कमी) को जानने में अधिक रुचिकर है बजाय पूर्व और पश्चात के व्यक्तिगत औसतों को जानने में ।

इसलिए इन परिणामों को उचित ढंग से दर्शाने के लिए हमें औसत अंतर के मानक विचलन की आवश्यकता है । इसकी गणना अंतर मानों पर वर्णनात्मक सांख्यिकी (टूल्स/डाटा एनालिसिस/डिसकिप्टिव स्टेस्टिक्स)की सहायता से की जा सकती है । सारांश सांख्यिकी तथा 95%प्रतिशत कानफिडेंस इन्टरवल ऑप्सन्स को चुनिये ।

परिणाम निम्नलिखित हैं :

स्तम्भ–1	
औसत	19.375
मानक त्रुटि	5.22677
माध्य	20.5
बहुलक	लागू नहीं
मानक विचलन	14.78356
प्रतिदर्श प्रसरण	218.5356
क्रुटोसिस	-0.57529
स्क्यूनैस	43
परास	-6
न्यूनतम	37
अधिकतम	155
योग	8
काउंट	–
कानफिडेंस स्तर (95%)	12.35936

यहां पर यह ध्यान देने योग्य है कि औसत को मानक त्रुटि से भाग करने पर वही मान प्राप्त होता है जो पिछली तालिकी की t-सांख्यिकी से प्राप्त हुई थी । अन्य महत्वपूर्ण सूचना 95 प्रतिशत कानफिडेंस इन्टरवल है । उपरोक्त तालिका से प्राप्त कानफिडेंस स्तर (95%)मान 12.259 है,कानफिडेंस इन्टरवल इस मान से थोड़ी अधिक अथवा थोड़ी कम औसत मान के बराबर है । अतः 95 प्रतिशत कानफिडेंस इन्टरवल पर औसत अंतर (7.01,31.74) है ।

इसे सही ढंग से इस प्रकार व्यक्त किया जा सकता है कि औसत भार हानि शून्य से अधिक है, द्वि पुच्छ p=0.008 इस बात का प्रमाण हैं कि आहार भार को कम करने में सक्षमहै। औसत भार हानि के आस पास95 प्रतिशत इन्टरवल(7.01,31.74)है ।

नोटः इस जांच को एकल पुच्छ जांच की तरह भी किया जा सकता है । इसके लिए एक्सेल तालिका से उचित t-सांख्यिकी तथा Pमान को प्रयोग करें ।

नोटः इस जांच को अंतर के शून्य के अतिरिक्त अन्य परिकाल्पित मानों के साथ भी किया जा सकता है । यद्यपि प्रायः शून्य मान को ही प्रयोग किया जाता है ।

एक्सैलयुगल t-जांच डायलाग बाक्स में अन्य परिकाल्पित मानों की प्रवष्ठि का अवसर प्रदान करती है ।

जैवमितिय विश्लेषणः

उदाहरण**1:** समंजन की शुष्टता के लिए काई–वर्ग के परीक्षण पर विचार करें जब डेटा की दो श्रेणियां हैं (अर्थात् k =2)। इस विश्लेषण वर्गों के किसी भी बड़ी संख्या के लिए आसानी से बढ़ाया जा सकता है। यहां $k = 4$

k **=2** के लिए काई–वर्ग समंजन की शुष्टता का परीक्षणः

H_0: प्रतिदर्श एक आबादी से 9:3:3:1के अनुपात मे पीली–चिकनी : पीली–झुर्रिया : हरी–चिकनी : हरी–झुर्रिया बीज आते हैं।

H_1: प्रतिदर्श एक आबादी से उपर्युक्त चार समलक्षणी के बीज 9:3:3:1 के अनुपात मे नहीं आते हैं।

प्रतिदर्श आंकणों प्रेक्षित मानों f_i के रूप मे, कोष्टक मे प्रत्याशित मानोंमानों F_i के साथ मे अभिलिखित किया गया

	पीली–चिकनी	पीली–झुर्रिया	हरी–चिकनी	हरी–झुर्रिया	n
f_i	152	39	53	6	250
(F_i)	140.625	46.875	46.875	15.625	

$v = k$-1=3

χ^2=8.972

0.025< P <0.05

अतः H_0 को अस्वीकार कर सकते हैं।

उदाहरण 2 द्विप्रतिदर्श t-जांच

H_0: $\mu_d \leq 5$

H_1: $\mu_d \geq 5$

भूखंड (j)	नए उर्वरक के साथ (X_{1j})	पुराने उर्वरक के साथ (X_{2j})	अन्तर (सेमी) d_j

1	67.4	60.6	6.8
2	72.8	66.6	6.2
3	68.4	64.9	3.5
4	66.0	61.8	4.2
5	70.8	61.7	9.1
6	69.6	67.2	2.4
7	67.2	62.4	4.8
8	68.9	61.3	7.6
9	62.6	56.7	5.9

N=9n d=5.611 bu/acre

$v=n-1$ s_d=0.701bu/ acre

t=d-5/0.701

=0.872

$T_{0.05(1),8}$=1.860

अतः H_0 को अस्वीकार नहीं कर सकते हैं।

आर सॉफ्टवयर (R-software)

समरेन्द्र दास
भा.कृ.अ.प.-भा.कृ.सां.अनु. संस्थान, नई दिल्ली-12
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
samarendra.das@icar.gov.in

R आँकड़ों और ग्राफिक्स के लिए एक उच्च-स्तरीय कंप्यूटर भाषा और वातावरण है। यह विभिन्न प्रकार के सरल और उन्नत सांख्यिकीय तरीके निष्पादित करता है और उच्च गुणवत्ता वाले ग्राफिक्स का उत्पादन करता है। इसके अलावा, आर एक कंप्यूटर भाषा है, इसलिए, हम नए कार्यों को लिख सकते हैं जो आर के उपयोग का विस्तार करते हैं। शुरुआत में रॉस इहाका और रॉबर्ट जेंटलमैन द्वारा सांख्यिकी विभाग, ऑकलैंड विश्वविद्यालय, ऑकलैंड, न्यूजीलैंड (इसलिए नाम) में लिखा गया था।) का है। R एक कमांड संचालित सांख्यिकीय पैकेज है, जिसे 17 प्रोग्रामर के "R Core Team" सहित कई योगदानकर्ताओं द्वारा बनाए रखा गया है, जो R स्रोत कोड (R Core Team, 2012) को संशोधित करने के लिए जिम्मेदार हैं।

पहली नजर में, यह उपयोग करने के लिए इसे कठिन बना सकता है। हालाँकि, इस कंप्यूटर प्रोग्राम का उपयोग करके आँकड़े सीखने के कई कारण हैं। दो सबसे महत्वपूर्ण हैं:

a) आर मुक्त है; आप इसे http://www.r-project.org से डाउनलोड कर सकते हैं और इसे अपने पसंद के किसी भी प्रकार के कंप्यूटर पर स्थापित कर सकते हैं।

b) आर आपको उन सभी सांख्यिकीय परीक्षणों को करने की अनुमति देता है जिनकी आपको आवश्यकता है, सरल से उच्च उन्नत वाले तक। इसका मतलब है कि आपको हमेशा अपने डेटा पर सही विश्लेषण करने में सक्षम होना चाहिए।

इसके अलावा, आर में उत्कृष्ट ग्राफिक्स और प्रोग्रामिंग क्षमताएं हैं, इसलिए इसका उपयोग शिक्षण और सीखने में सहायता के रूप में किया जा सकता है। R की ताकत यह है कि सांख्यिकीय विश्लेषणों के साथ-साथ अच्छी तरह से डिज़ाइन किए गए प्रकाशन-गुणवत्ता वाले ग्राफिक्स का उत्पादन किया जा सकता है। आर सभी ऑपरेटिंग सिस्टम (लिनक्स, मैक और विंडोज) पर चलता है।

R डाउनलोड और पर्यावरण

R, CRAN मिरर साइटों (CRAN: व्यापक R संग्रह नेटवर्क) के नेटवर्क से आसानी से उपलब्ध है। R डाउनलोड करने और इंस्टॉल करने के लिए www.r-project.org पर जाएं और पास में एक CRAN मिरर चुनें। R एक कंसोल के माध्यम से संचालित कोड काम करता है, न कि उन मेनू के साथ जिनका उपयोग आप अन्य सॉफ़्टवेयर से कर सकते हैं। आर-कंसोल सिर्फ एक कैलकुलेटर है। अपने विश्लेषण के चरणों का दस्तावेजीकरण करने के लिए, आप अपने आर कोड को एक टेक्स्ट एडिटर में लिखेंगे (कोड के छोटे बिट्स को छोड़कर जिन्हें आपको सहेजने की आवश्यकता नहीं है)। पाठ संपादक से, आप कॉपी या भेज सकते हैं (यदि आपका संपादक आर के साथ बातचीत करता है) फ़ंक्शन कॉल को निष्पादित करने के लिए आर कंसोल को कोड। आप आर द्वारा उत्पादित परिणामों को पाठ फ़ाइलों में सहेज सकते हैं या विभिन्न प्रारूपों में ग्राफिक्स का उत्पादन कर सकते हैं। जब आप अपना R सत्र बंद करते हैं, तो R-कंसोल स्वयं सामान्य रूप से सहेजा नहीं जाता है। हालांकि, किसी भी समय अपने विश्लेषण को फिर से संगठित करने में सक्षम होने के लिए, आपको अपने आर कोड वाले टेक्स्ट फ़ाइल (फाइलों) को सहेजना चाहिए। यद्यपि आप आर कोड को लिखने और सहेजने के लिए किसी भी टेक्स्ट एडिटर का उपयोग कर सकते हैं (जैसे नोटपैड), यह एक टेक्स्ट एडिटर स्थापित करने की सिफारिश की गई है जो आर भाषा को पहचानता है, जैसे कि टिन-आर (http://www.sciviews.org/Tinn- R), RStudio (www.rstudio.org), या Emacs।

अपने कंप्यूटर पर आर स्थापित करने के बाद, यदि स्थापित नहीं है, तो http://www.r-project.org से मुफ्त में नवीनतम संस्करण डाउनलोड करें और आधार प्रणाली स्थापित करें। आपको अभी तक कोई अतिरिक्त पैकेज स्थापित करने की

आवश्यकता नहीं है। एक बार जब आप इसे स्थापित कर लेते हैं, तो इसे शुरू करें और आपको कुछ इस तरह प्रस्तुत करना चाहिए:

R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"

Copyright (C) 2016 The R Foundation for Statistical Computing

Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.

You are welcome to redistribute it under certain conditions.

Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.

Type 'contributors()' for more information and

'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or

'help.start()' for an HTML browser interface to help.

Type 'q()' to quit R.

R में एक संपूर्ण सहायता "help.start ()" टाइप करके देखी जा सकती है। R में खोज इंजन जिसके बाद "कीवर्ड बाय टॉपिक" की सूची उपलब्ध है और इसे देखा जा सकता है।

आर की मूल बातें

इनपुट

यहां हम पता लगाते हैं कि आर सत्र में डेटा सेट को कैसे परिभाषित किया जाए। केवल दो आदेशों का पता लगाया जाता है। पहला डेटा के सरल असाइनमेंट के लिए है, और दूसरा डेटा फ़ाइल में पढ़ने के लिए है। आर सत्र में डेटा पढ़ने के कई तरीके हैं, लेकिन हम इसे सरल रखने के लिए सिर्फ दो पर ध्यान केंद्रित करते हैं।

संख्याओं की सूची को संग्रहीत करने का सबसे सीधा आगे तरीका सी कमांड का उपयोग करके असाइनमेंट के माध्यम से है। " सी कमांड के साथ एक सूची निर्दिष्ट की गई है, और असाइनमेंट "<-" प्रतीकों के साथ निर्दिष्ट किया गया है। संख्याओं की सूची का वर्णन करने के लिए उपयोग किया जाने वाला एक और शब्द इसे "वेक्टर" कहना है। सी कमांड के भीतर संख्याओं को कॉमा द्वारा अलग किया जाता है।

उदाहरण के लिए, हम "a" नामक एक नया चर बना सकते हैं जिसमें 3, 5, 7 और 9 नंबर होंगे:

```
> a <- c(3,5,7,9)
```

जब आप इस कमांड को दर्ज करते हैं तो आपको नई कमांड लाइन को छोड़कर कोई आउटपुट नहीं देखना चाहिए। कमांड संख्या की एक सूची बनाता है जिसे "ए" कहा जाता है। यह देखने के लिए कि "a" में कौन सी संख्याएँ शामिल हैं, बस "a" टाइप करें और एंटर की दबाएं और परिणाम होगा:

```
> a

[1] 3 5 7 9
```

यदि आप संख्याओं में से किसी एक के साथ काम करना चाहते हैं, तो आप चर का उपयोग करके इसे प्राप्त कर सकते हैं और फिर वर्ग कोष्ठक जो यह दर्शाता है कि कौन सी संख्या:

```
> a[2]

[1] 5
```

1.2। डेटा फ़ाइल पढ़ना

दुर्भाग्य से, यह केवल कुछ डेटा बिंदुओं के लिए दुर्लभ है, जिन्हें आपको प्रॉम्प्ट पर टाइप करने में कोई आपत्ति नहीं है। जटिल संबंधों (जैसे जीनोमिक डेटा) के साथ बहुत अधिक डेटा बिंदु होना बहुत आम है। यहां हम यह जांचेंगे कि रीड.टेबल और अन्य फ़ंक्शन का उपयोग करके किसी फ़ाइल से डेटा सेट कैसे पढ़ें, लेकिन पहले डेटा फ़ाइल कैसे बनाएं।

data.frame फ़ंक्शन डेटा फ़्रेम बनाता है, चर के कसकर युग्मित संग्रह जो आर के अधिकांश मॉडलिंग सॉफ़्टवेयर द्वारा मूलभूत डेटा संरचना के रूप में उपयोग किए जाने वाले मैट्रिसेस और सूचियों के कई गुणों को साझा करता है।

```
data.frame(..., row.names = NULL, check.rows = FALSE,
      check.names = TRUE,
      stringsAsFactors = default.stringsAsFactors())
default.stringsAsFactors()
```

read.table तालिका प्रारूप में एक फ़ाइल पढ़ता है और फ़ाइल में फ़ील्ड के लिए लाइनों और चर के अनुरूप मामलों के साथ, इससे एक डेटा फ़्रेम बनाता है।

```
read.table(file, header = FALSE, sep = "", quote = "\"'",
```

```
dec = ".", row.names, col.names,
as.is = !stringsAsFactors,
na.strings = "NA", colClasses = NA, nrows = -1,
skip = 0, check.names = TRUE, fill = !blank.lines.skip,
strip.white = FALSE, blank.lines.skip = TRUE,
comment.char = "#",
allowEscapes = FALSE, flush = FALSE,
stringsAsFactors = default.stringsAsFactors(),
encoding = "unknown")
```

लिखना: डेटा (आमतौर पर एक मैट्रिक्स) x फ़ाइल फ़ाइल के लिए लिखा जाता है। यदि x एक द्वि-आयामी मैट्रिक्स है, तो इसे आंतरिक प्रतिनिधित्व के रूप में फ़ाइल में कॉलम प्राप्त करने के लिए इसे स्थानांतरित करने की आवश्यकता हो सकती है।

```
write (x, file = "data", ncolumns = if(is.character(x)) 1 else 5, append = FALSE, sep = " ")
```

X **डेटा बाहर लिखा जाना है**

File **एक कनेक्शन, या एक चरित्र स्ट्रिंग को लिखने के लिए फ़ाइल का नामकरण। यदि** "", **मानक आउटपुट कनेक्शन पर प्रिंट करें।**

ncolumns **डेटा लिखने के लिए कॉलम की संख्या।**

append **यदि** TRUE **डेटा** x **को कनेक्शन से जोड़ा जाता है।**

Sep **स्तंभों को अलग करने के लिए प्रयुक्त एक स्ट्रिंग।** Sep = "\ t" **का उपयोग टैब सीमांकित आउटपुट देता है**; **डिफ़ॉल्ट** "" **है।**

लिखने में सक्षम। अपने आवश्यक तर्क एक्स को प्रिंट करता है (एक फ़ाइल या कनेक्शन के लिए यह एक डेटा फ्रेम में बदलने के बाद अगर यह एक और न ही मैट्रिक्स है)।

```
write.table(x, file = "", append = FALSE, quote = TRUE, sep = " ",
        eol = "\n", na = "NA", dec = ".", row.names = TRUE,
        col.names = TRUE, qmethod = c("escape", "double"))
```

2. बुनियादी डेटा प्रकार

2.1। चर प्रकार

नंबर

वास्तविक संख्याओं के साथ काम करने का तरीका पहले से ही पहले अध्याय में शामिल किया गया है और यहां संक्षेप में चर्चा की गई है। किसी संख्या को संग्रहीत करने का सबसे मूल तरीका एक संख्या का असाइनमेंट बनाना है:

```
a <- 3
```

"<-" आर को प्रतीक के दाईं ओर संख्या लेने और एक चर में संग्रहीत करने के लिए कहता है जिसका नाम बाईं ओर दिया गया है। आप "=" प्रतीक का भी उपयोग कर सकते हैं। जब आप एक असाइनमेंट बनाते हैं तो आर किसी भी जानकारी को प्रिंट नहीं करता है।

यदि आप यह देखना चाहते हैं कि किसी चर का मान किसी रेखा पर चर का नाम किस प्रकार है और एंटर की दबाएं:

```
> a

[1] 3
```

यह आपको सभी प्रकार के बुनियादी कार्यों को करने और संख्याओं को बचाने की अनुमति देता है:

```
> b <- sqrt(a*a+3)

> b

[1] 3.464102
```

स्ट्रिंग्स

आप केवल स्टोरिंग नंबर तक सीमित नहीं हैं। आप स्ट्रिंग्स को स्टोर भी कर सकते हैं। उद्धरण का उपयोग करके एक स्ट्रिंग निर्दिष्ट की जाती है। दोनों सिंगल और डबल कोट्स काम करेंगे:

```
> a <- "hello"

> a

[1] "hello"

> b <- c("hello","there")

> b

[1] "hello" "there"

> b[1]

[1] "hello"

> typeof(a)

[1] "character"

> a = character(20)

> a "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" "" ""
```

कारकों

एक और महत्वपूर्ण तरीका, आर डेटा को एक कारक के रूप में संग्रहीत कर सकता है। अक्सर एक प्रयोग में कुछ व्याख्यात्मक चर के विभिन्न स्तरों के लिए परीक्षण शामिल होते हैं। उदाहरण के लिए, जब एक पेड़ की वृद्धि दर पर कार्बन डाइऑक्साइड के प्रभाव को देखते हुए आप यह देखने की कोशिश कर सकते हैं कि कार्बन डाइऑक्साइड के अलग-अलग पूर्व निर्धारित सांद्रता के संपर्क में आने पर विभिन्न पेड़ कैसे बढ़ते हैं। विभिन्न स्तरों को कारक भी कहा जाता है।

नीचे दिए गए उदाहरण (उदाहरण के लिए पेड़ डेटा फ़ाइल) के लिए, फ़ाइल में कई चर कारक हैं:

```
> summary(tree$CHBR)

A1  A2  A3  A4  A5  A6  A7  B1  B2  B3  B4  B5  B6  B7  C1  C2  C3  C4  C5  C6

3   1   1   3   1   3   1   1   3   3   3   3   3   3   1   3   2   3   1   1

C7 CL6 CL7  D1  D2  D3  D4  D5  D6  D7

1   1   1   1   1   3   1   1   1   4
```

इस डेटा सेट में कई स्तंभ कारक हैं, लेकिन शोधकर्ताओं ने विभिन्न स्तरों को इंगित करने के लिए संख्याओं का उपयोग किया। उदाहरण के लिए, "A1" लेबल वाला पहला कॉलम एक कारक है। प्रत्येक पेड़ एक ऐसे वातावरण में उगाया गया था जिसमें कार्बन डाइऑक्साइड के चार अलग-अलग संभावित स्तरों में से एक था। शोधकर्ताओं ने काफी समझदारी से इन चार वातावरणों को 1, 2, 3 और 4 के रूप में लेबल किया। दुर्भाग्य से, आर यह निर्धारित नहीं कर सकता है कि ये कारक हैं और उन्हें यह मान लेना चाहिए कि वे नियमित संख्या हैं।

3. बुनियादी संचालन और संख्यात्मक विवरण

हम कुछ बुनियादी ऑपरेशनों को देखते हैं जिन्हें आप संख्याओं की सूची पर कर सकते हैं। यह माना जाता है कि आप डेटा दर्ज करना जानते हैं या डेटा फ़ाइलों को पढ़ना चाहते हैं जो उपरोक्त अनुभाग में शामिल हैं और आपको मूल डेटा प्रकारों के बारे में पता है।

3.1। बुनियादी संचालन

एक बार जब आपके पास एक वेक्टर (या संख्याओं की एक सूची) स्मृति में सबसे बुनियादी संचालन उपलब्ध हैं। अधिकांश बुनियादी ऑपरेशन एक पूरे वेक्टर पर कार्य करेंगे और एक ही आदेश के साथ बड़ी संख्या में गणना करने के लिए जल्दी से उपयोग किए जा सकते हैं। ध्यान देने वाली एक बात है, यदि आप एक से अधिक वेक्टर पर एक ऑपरेशन करते हैं तो अक्सर यह आवश्यक होता है कि वैक्टर सभी में समान संख्या में प्रविष्टियाँ हों।

मूल उदाहरण

R कोड लाइन को लाइन से चलाता है। यही है, आप इसे एक बात बताते हैं, और यह इसे तुरंत करता है। (कभी-कभी यदि कोड की हमारी एक "लाइन" सुपर लंबी होती है, तो यह वास्तव में एक पृष्ठ पर कई लाइनों के रूप में लिखा जाएगा, लेकिन आर इसे कोड के एक सुपर-लॉन्ग वाक्य के रूप में मानता है)।

संख्याओं के साथ, हम कैलकुलेटर की तरह आर का उपयोग कर सकते हैं। जब हम 3 + 7 टाइप करते हैं और एंटर करते हैं, तो कंसोल विंडो में जो दिखाई देता है, उसका एक उदाहरण निम्नलिखित है।

```
> 3+7
[1] 10
```

Basic arithmetic operators	code	Results
+	3+7	3+7=10

-	3-7	3-7=-4
*	3*7	3X7=21
/	3/7	3/7=0.4286
sqrt	sqrt(3)	√3=1.732045
log	log(2)	Natural Logarithm
exp	exp(log(2))	2
sin, cos, tan	sin(a), cos(a), tan(a)	sin(a), cos(a), tan(a)
$\sin^{-1}$, $\cos^{-1}$, $\tan^{-1}$	asin(a),a cos(a),atan(a)	asin(a),a cos(a),atan(a)

हम नामों का उपयोग करके वस्तुओं को संग्रहीत भी कर सकते हैं। हम इस वर्ग में नामांकित डेटा फ़्रेम के साथ सबसे अधिक बार देखते हैं। (उर्फ डेटा सेट)। हम तालिकाओं, फ़ंक्शन आउटपुट या एकल मान भी संग्रहीत करेंगे। एक सरल उदाहरण निम्नलिखित कोड है:

se <- sqrt(.75*.25/200)

उदाहरण के लिए: मैं अपने कार्यक्षेत्र में "से" के रूप में 200 टिप्पणियों के साथ .75 के नमूने अनुपात के लिए मानक त्रुटि को संग्रहीत करना चाहता हूं। यह सुविधाजनक है अगर मैं इसे समीकरणों में बार-बार उपयोग करने जा रहा हूं। आप देखेंगे कि यदि आप कोड की इस लाइन को चलाते हैं, तो आपके कंसोल में कोई आउटपुट दिखाई नहीं देता है। लेकिन आपके कार्यक्षेत्र में एक नया "मूल्य" प्रकट होता है, जिसे se कहा जाता है। आप "<-" के बजाय मान निर्दिष्ट करने के लिए "=" का उपयोग कर सकते हैं। पाठ्यपुस्तक "=" का उपयोग करती है, लेकिन कई एक सम्मेलन के रूप में तीर का उपयोग करना पसंद करते हैं; जैसा कि आप अधिक कोड लिखते हैं, आप अपनी शैली विकसित करेंगे।

Note that R is case sensitive. The object se is not the same as SE.

इसके अलावा, आर सॉफ्टवेयर अनुसंधान के सभी क्षेत्रों से प्राप्त अधिकांश प्रयोगात्मक डेटा का विश्लेषण करता है। विवरणात्मक सांख्यिकी, प्रतिगमन, सहसंबंध, रैखिक मॉडल, विचरण का विश्लेषण, पूरी तरह से यादृच्छिक डिजाइन, यादृच्छिक पूर्ण ब्लॉक डिजाइन, लैटिन वर्ग डिजाइन, प्रमुख घटक विश्लेषण, क्लस्टर विश्लेषण, आदि जैसे सांख्यिकीय विश्लेषण आर। विश्लेषण में उपलब्ध हैं जो उपरोक्त तकनीकों पर आधारित हैं। व्यावहारिक रूप से, विस्तार से वास्तविक जीवन के उदाहरणों से निपटा जाता है।

इन विश्लेषणों के बारे में मदद आसानी से help.start () को शेल प्रॉम्प्ट पर टाइप करके और संक्षेप में दी जा सकती है:

कक्षाएं: डेटा प्रकार

ओ एनए: गुम मान

ओ श्रेणी: श्रेणीबद्ध डेटा

ओ चरित्र: चरित्र डेटा ("स्ट्रिंग") संचालन

o जटिल: जटिल संख्या

• डेटा: वातावरण, स्कोपिंग, पैकेज

• डेटासेट: डेटा द्वारा उपलब्ध डेटासेट ()

• सूची: सूचियाँ

• हेरफेर: डेटा हेरफेर

• पैकेज: पैकेज सारांश

• sysdata: बुनियादी प्रणाली चर

ग्राफिक्स

• aplot: मौजूदा प्लॉट / आंतरिक भूखंड में जोड़ें

• रंग: रंग, पट्टियाँ आदि

• डिवाइस: ग्राफिकल डिवाइस

• dplot: प्लॉटिंग से संबंधित संगणना

• गतिशील: गतिशील ग्राफिक्स

• hplot: उच्च-स्तरीय भूखंड

• ipl: प्लॉट के साथ बातचीत

MASS (पुस्तक) का उपयोग करता है

• वर्गीकरण: वर्गीकरण

• तंत्रिका: तंत्रिका नेटवर्क

• स्थानिक: स्थानिक सांख्यिकी

गणित

• एरीथ: बेसिक अंकगणित और छंटनी

• सरणी: मैट्रिसेस और एरेस

o बीजगणित: रैखिक बीजगणित

• रेखांकन: रेखांकन (ग्राफिक्स नहीं), यानी नोड्स

प्रोग्रामिंग, इनपुट / Ouput, और विविध

• IO: इनपुट / आउटपुट

ओ कनेक्शन: इनपुट / आउटपुट कनेक्शन

ओ डेटाबेस: डेटाबेस के लिए इंटरफेस

ओ फाइल: इनपुट / आउटपुट फाइलें

• डीबगिंग: डीबगिंग टूल

• प्रलेखन: प्रलेखन

• पर्यावरण: सत्र पर्यावरण

• त्रुटि: त्रुटि हैंडलिंग

• आंतरिक: आंतरिक ऑब्जेक्ट (एपीआई का हिस्सा नहीं)

• पुनरावृति: लूपिंग और पुनरावृत्ति

• तरीके: तरीके और सामान्य कार्य

• विविध: विविध

आंकड़े

• क्लस्टर: क्लस्टरिंग

डेटा सेट उत्पन्न करने के लिए कार्य

• डिजाइन: डिजाइन प्रयोगों

• वितरण: संभाव्यता वितरण और यादृच्छिक संख्या

• htest: सांख्यिकीय इंजेक्शन

• मॉडल: सांख्यिकीय मॉडल

ओ प्रतिगमन: प्रतिगमन

 नॉनलाइनर: गैर-रेखीय प्रतिगमन

• बहुभिन्नरूपी: बहुभिन्नरूपी तकनीक

• नॉनपैरेमेट्रिक: नॉनपैरेमेट्रिक सांख्यिकी

• मजबूत: मजबूत / प्रतिरोधी तकनीक

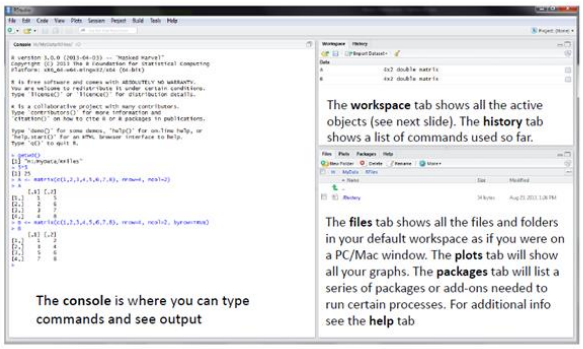
• चिकनी: वक्र (और सतह) चौरसाई

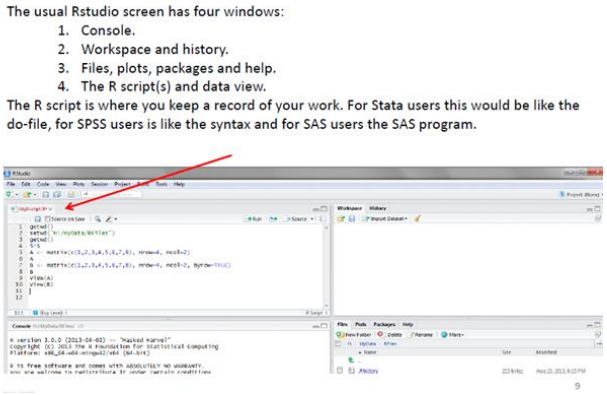
ओ loess: ढीली वस्तुओं

RStudio

RStudio सांख्यिकीय प्रोग्रामिंग सॉफ़्टवेयर R के लिए एक उपयोगकर्ता इंटरफ़ेस है। जबकि कुछ ऑपरेशन माउस से इंगित और क्लिक करके किए जा सकते हैं, प्रोग्राम कोड लिखना सीखना आवश्यक है। यह एक नई भाषा सीखने की तरह है - विशिष्ट वाक्यविन्यास, व्याकरण और शब्दावली है, और इसका उपयोग करने में समय लगेगा। आर स्टूडियो सीखना अंततः आर पर डेटा का विश्लेषण और कल्पना करते समय पूर्ण नियंत्रण, लचीलापन और रचनात्मकता देगा, लेकिन इस नई भाषा में प्रवाह में समय लगेगा।

एक बार जब आप R डाउनलोड कर लेते हैं, तो आप http://www.rstudio.com/ से RStudio स्थापित कर सकते हैं, "अभी डाउनलोड करें" पर क्लिक करके, और फिर "RStudio डेस्कटॉप डाउनलोड करें" पर क्लिक करें। अपने ऑपरेटिंग सिस्टम के लिए उपयुक्त संस्करण का चयन करें और डाउनलोड करें। जब आप RStudio खोलेंगे तो आपको निम्न स्क्रीन दिखाई देगी और चार विंडो होंगी:





चार खिड़कियों के रूप में वर्णित किया जा सकता है:

लिपि

स्क्रिप्ट R कमांड की एक सूची को संग्रहीत करने के लिए एक दस्तावेज़ है। जब आप पहली बार RStudio खोलते हैं तो यह विंडो प्रकट नहीं हो सकती है। एक नई स्क्रिप्ट बनाने के लिए, "फाइल -> नया -> आर स्क्रिप्ट" पर क्लिक करें।

कंसोल

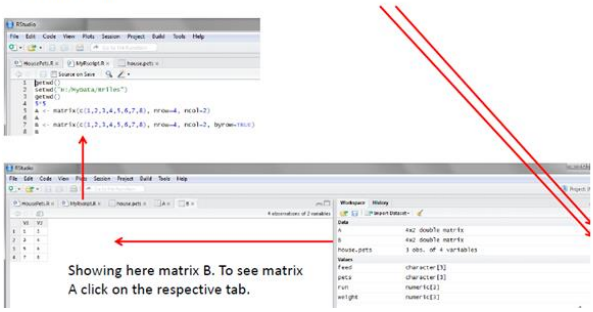
यहां आउटपुट दिखाई देता है। > संकेत (जिसे "प्रॉम्प्ट" भी कहा जाता है) का अर्थ है कि आर आज्ञाओं को स्वीकार करने के लिए तैयार है। आप कमांड को सीधे कंसोल में टाइप कर सकते हैं। हालाँकि, इसके बजाय स्क्रिप्ट विंडो में टाइप करना और वहाँ से कमांड चलाना एक अच्छी आदत है। कंसोल में कुछ भी नहीं बचाया जा सकता है। हालाँकि आप अपने आदेश को स्क्रिप्ट फ़ाइल में सहेज सकते हैं, और फिर बाद में अपने विश्लेषण को दोहरा सकते हैं। यदि आप किसी बड़ी परियोजना पर काम कर रहे हैं या आप बाद में वापस आने के लिए अपना कोड रखना चाहते हैं तो यह विशेष रूप से सहायक है।

कार्यस्थान

यह कार्यस्थान विंडो आपके पास वर्तमान में उपलब्ध वस्तुओं को सूचीबद्ध करती है। फ़ंक्शंस जो "बेस आर" या पैकेज का हिस्सा हैं, वे यहां दिखाई नहीं देंगे (उस प्रैक्टिकल को बनाने के लिए बस बहुत सारे हैं!) विशेष फ़ंक्शन जो आप खुद लिखते हैं या जो हैं।

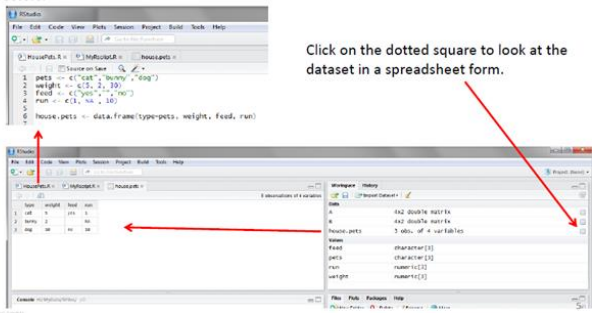
Workspace tab (1)

The workspace tab stores any object, value, function or anything you create during your R session. In the example below, if you click on the dotted squares you can see the data on a screen to the left.



Workspace tab (2)

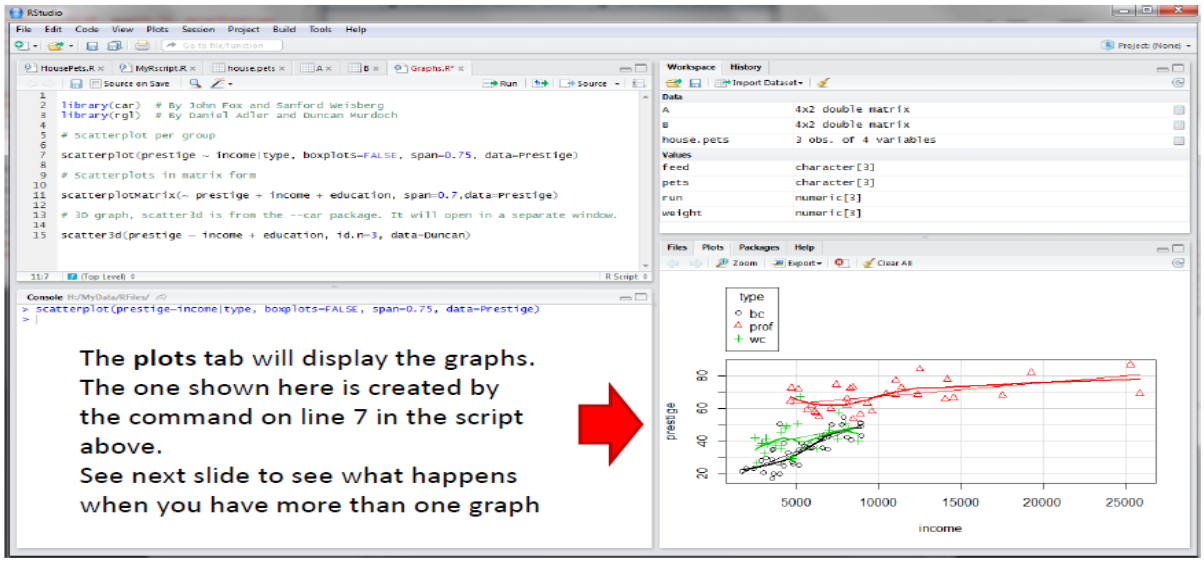
Here is another example on how the workspace looks like when more objects are added. Notice that the data frame `house.pets` is formed from different individual values or vectors.



प्लॉट / सहायता

अंतिम विंडो में कई टैब हैं, जिसमें एक खोज सुविधा के साथ एक सहायता टैब भी शामिल है। जब आप भूखंड बनाते हैं तो वे इस विंडो में दिखाई देंगे, जिसे आप बेहतर दृश्य प्राप्त करने के लिए आकार बदल सकते हैं। "फ़ाइलें" टैब आपको आपके द्वारा पहले लिखी गई आर लिपियों तक पहुंचने के एक तरीके के रूप में आपके कंप्यूटर पर फाइलें भी दिखाता है। सावधान रहें- इस विंडो में फ़ाइलों को हटाने से उन्हें आपके कंप्यूटर से हटा दिया जाता है। आर स्टूडियो में प्लॉट विंडो के रूप में कल्पना की जा सकती है:

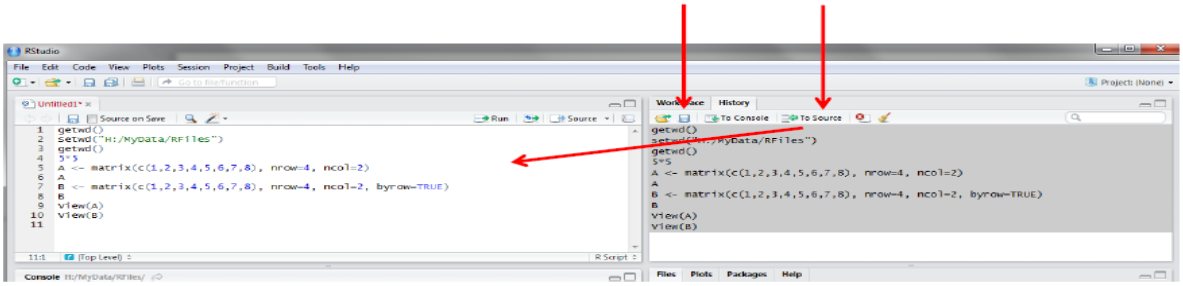
Plots tab (1)



इसके अलावा, इतिहास की खिड़की को अच्छी तरह से देखा जा सकता है:

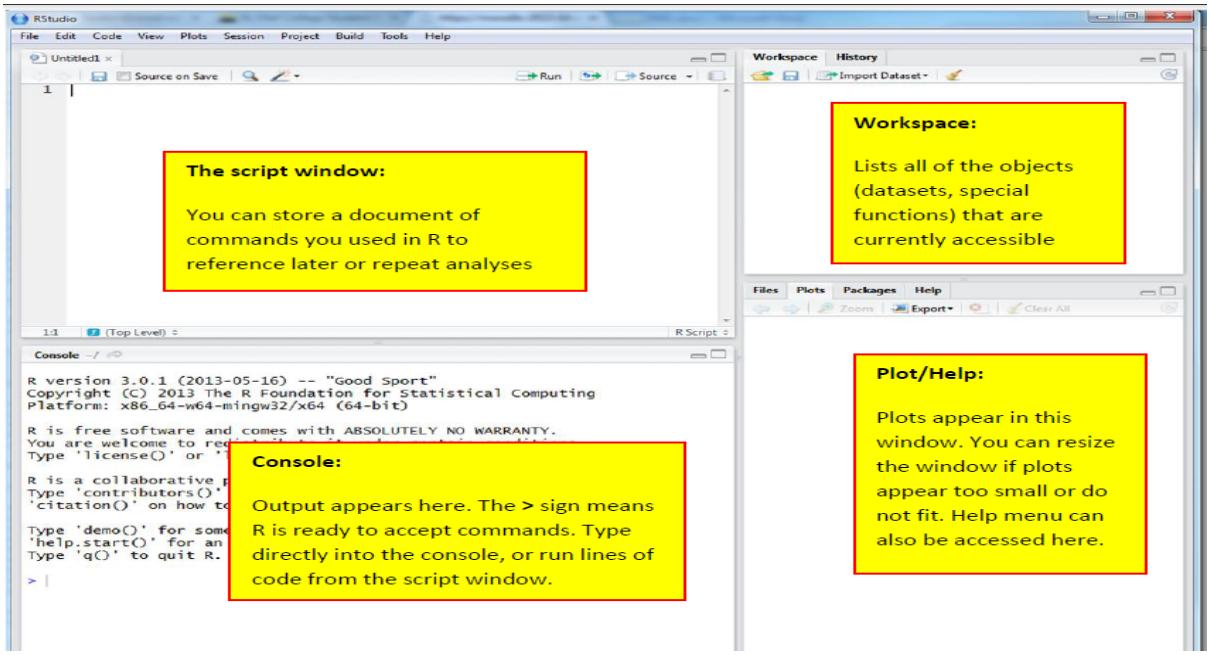
History tab

The history tab keeps a record of all previous commands. It helps when testing and running processes. Here you can either **save** the whole list or you can **select** the commands you want and send them to an R script to keep track of your work.
In this example, we select all and click on the "To Source" icon, a window on the left will open with the list of commands. Make sure to save the 'untitled1' file as an *.R script.



6

हालाँकि, R स्टूडियो की इन चार खिड़कियों को अच्छी तरह से देखा जा सकता है:

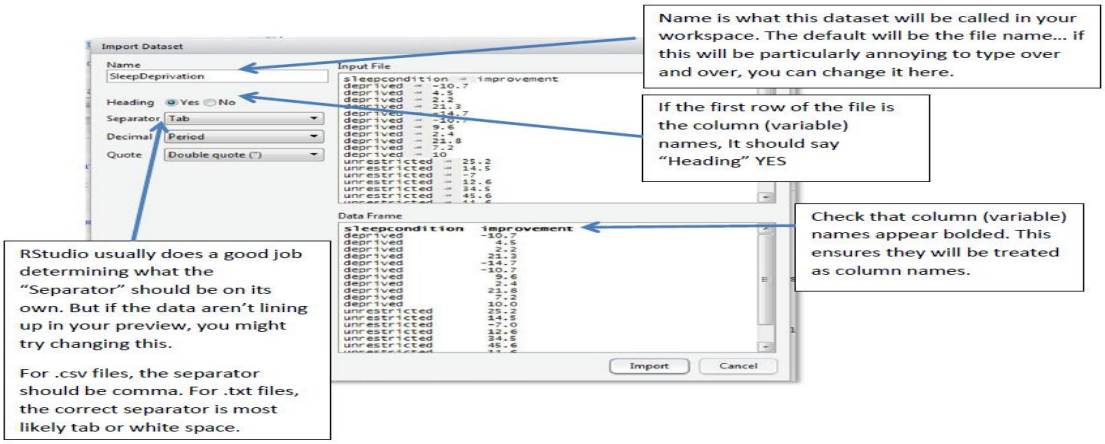


RStudio में कार्यस्थान और डेटासेट लोड हो रहे हैं

.RData फ़ाइल एक्सटेंशन का उपयोग करके कार्यस्थानों को सहेजा जाता है। एक कार्यक्षेत्र कई डेटासेट या आपके द्वारा लिखे गए कार्यों के एक सेट को स्टोर करने का एक सुविधाजनक तरीका है, खासकर जब कोड को चलाने के लिए डेटासेट का उत्पादन करने में लंबा समय लग सकता है। कार्यस्थान लोड करने के लिए, ऊपरी दाएँ RStudio विंडो में "कार्यस्थान" टैब के अंतर्गत फ़ोल्डर आइकन पर क्लिक करें। जहाँ भी आपने कार्यक्षेत्र को सहेजा है और उसे खोलें, पर नेविगेट करें। अब आपको कार्यक्षेत्र में वस्तुओं की एक सूची देखनी चाहिए । अपने कार्यक्षेत्र में डेटासेट लोड करने के लिए, आपको आयात डेटा बटन पर क्लिक करने

और "फ़ाइल से" या "URL से" उपयुक्त के रूप में चयन करने की आवश्यकता है। आप csv या txt फ़ाइलों को लोड कर सकते हैं जिन्हें आपने "फ़ाइल से" अपने कंप्यूटर पर सहेजा है। जब डेटा ऑनलाइन पाठ फ़ाइलों के रूप में दिखाई देते हैं, तो आप उन्हें सीधे URL से लोड करने में सक्षम हो सकते हैं।

अब आपको केवल यह सुनिश्चित करने की आवश्यकता है कि डेटा फ़्रेम का पूर्वावलोकन सही है या नहीं और जैसा दिखाया गया है:

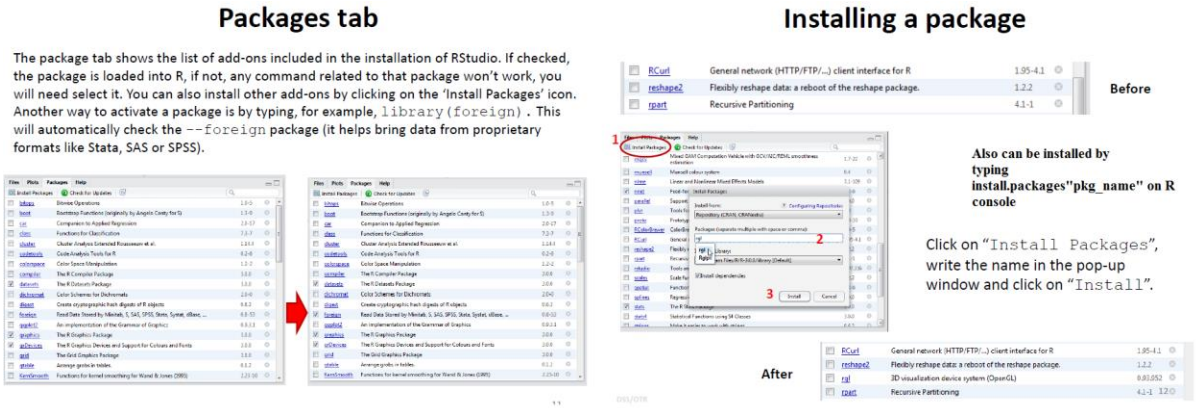


Once you import the dataset, a new data frame will appear in your workspace with whatever name was in the "Name" box.

संकुल

जबकि कई उपयोगी फ़ंक्शंस "बेस आर" में शामिल हैं, उपयोगकर्ता और डेवलपर्स विशेष कार्यों और डेटासेट के साथ अपने स्वयं के ऐड-ऑन पैकेज बना और जमा कर सकते हैं। इन पैकेजों तक पहुँचने के लिए दो चरणों की आवश्यकता होती है: पैकेज को अपने कंप्यूटर पर स्थापित करना (केवल एक बार करने की आवश्यकता होती है) और अपने कार्यक्षेत्र में पुस्तकालय को लोड करने की आवश्यकता है (हर बार जब आप RStudio खोलते हैं तो ऐसा करने की आवश्यकता होती है)।

संकुल बिंदु द्वारा स्थापित किया जा सकता है और RStudio में क्लिक कर सकते हैं।



कार्यशील निर्देशिका बदलना

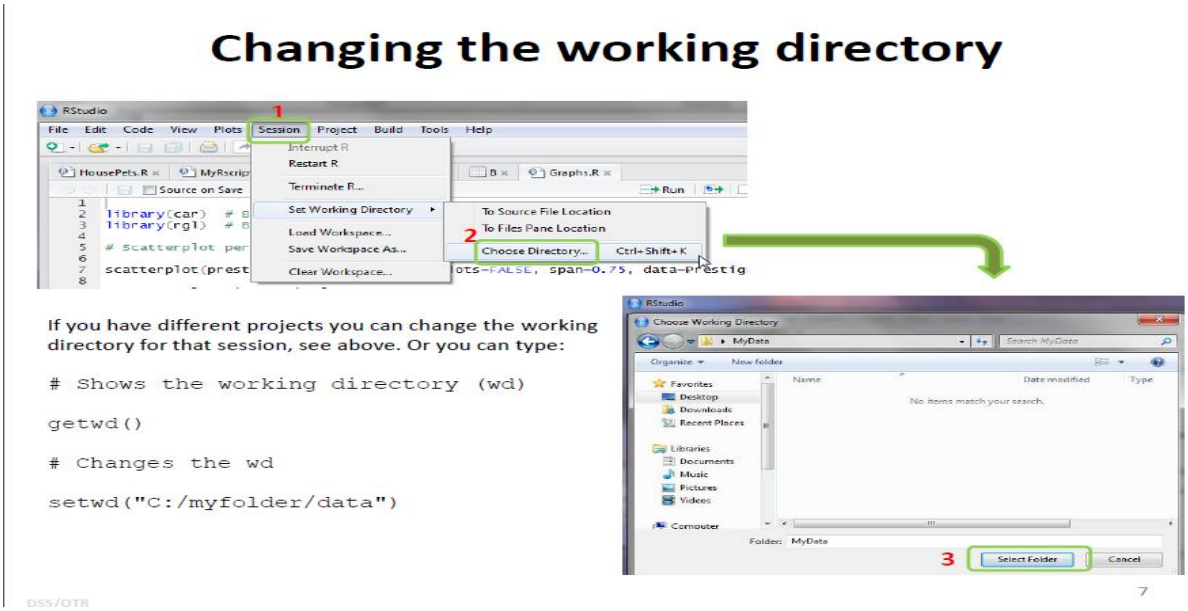
R को हमेशा कंप्यूटर पर एक निर्देशिका में इंगित किया जाता है। यह आसानी से पता लगाया जा सकता है कि कौन सी डायरेक्टरी गेटवे (वर्किंग डायरेक्टरी प्राप्त करें) फ़ंक्शन को चलाकर; इस फ़ंक्शन का कोई तर्क नहीं है। कार्यशील निर्देशिका को

बदलने के लिए, सेटवार्ड का उपयोग करें और वांछित फ़ोल्डर में पथ निर्दिष्ट करें। dir - एक कार्यशील निर्देशिका निर्दिष्ट करें। इसके अलावा, getwd R प्रक्रिया की वर्तमान कार्यशील निर्देशिका का प्रतिनिधित्व करते हुए एक निरपेक्ष फ़ाइलपथ लौटाता है; setwd (dir) का उपयोग कार्य निर्देशिका को dir में सेट करने के लिए किया जाता है।

Usage

getwd()

setwd(dir)



स्क्रिप्ट विंडो में कोड लिखना

यह वास्तव में सीधे सांत्वना में सब कुछ टाइप करने के लिए आकर्षक हो सकता है- और यदि आप केवल एक या दो लाइनों का विश्लेषण कर रहे हैं जिसे आप कभी नहीं दोहराएंगे, तो यह ठीक हो सकता है। हालांकि, होमवर्क और प्रोजेक्ट्स करते समय आपके द्वारा चलाए गए कोड की एक प्रति होना आवश्यक होगा। मैं अक्सर आपके साथ R स्क्रिप्ट साझा करूंगा जिसमें उदाहरण शामिल हैं। इन्हें रखना एक अच्छा विचार है, और यहां तक कि अपनी टिप्पणी और नोट्स भी जोड़ें क्योंकि हम उन्हें कक्षा में उपयोग करते हैं।

स्क्रिप्ट विंडो में कोड लिखने के लाभ:

1. ऑटो-रंग और कोष्ठक हाइलाइटिंग त्रुटियों को खोजने में आसान बनाते हैं।

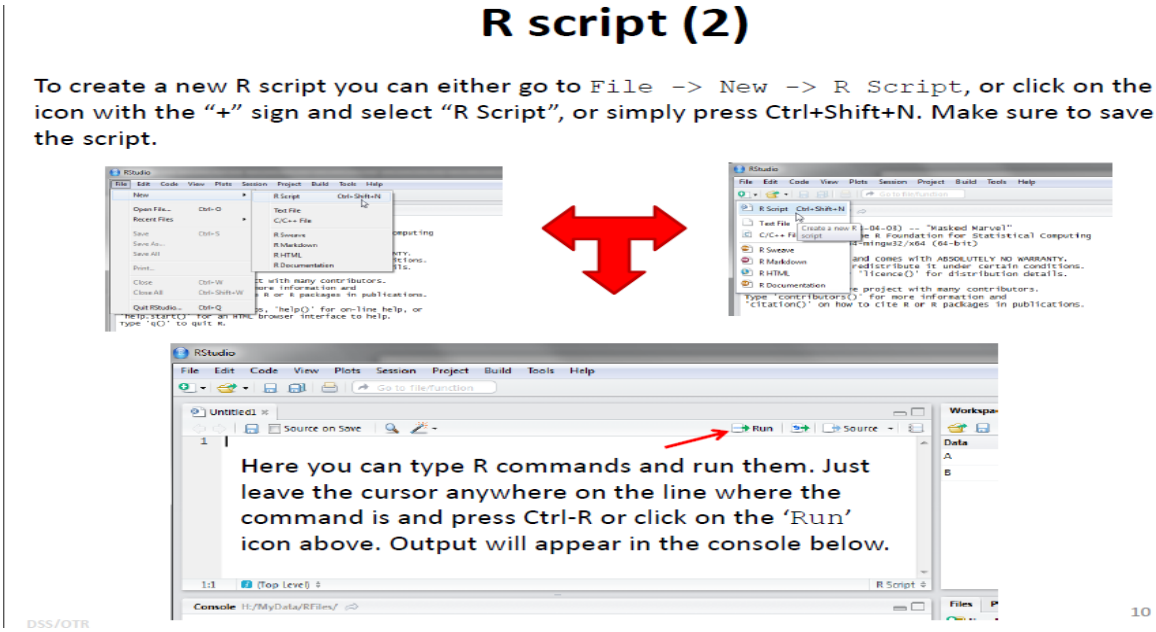
2. आप अपने कोड को सहेज सकते हैं और बाद में संदर्भ के लिए खुद को नोट्स लिख सकते हैं।

3. प्रोजेक्ट्स पर काम करते समय, या प्रश्न होने पर मेरे साथ साझा करने के लिए सहपाठियों के साथ अपना कोड साझा करना आसान बनाता है।

4. अपने विश्लेषण को दोहराने योग्य बनाता है, संपादित करने और कॉपी करने में आसान।

लिपियों में फ़ाइल एक्सटेंशन ".R" होता है, यदि आपने अपने कंप्यूटर पर R स्थापित नहीं किया है, तो आप एक। संपादक जैसे नोटपैड (हालांकि तब आप केवल सादे पाठ, रंगों को नहीं देख सकते हैं) का उपयोग करके .R फाइलें देख सकते हैं। सादा पाठ फ़ाइलें (.txt) भी स्क्रिप्ट फ़ाइलों के रूप में RStudio में खोली जा सकती हैं।

स्क्रिप्ट फ़ाइलों के बारे में महान चीजों में से एक टिप्पणी शामिल करने की क्षमता है। ये R कमांड के साथ डाले गए नोट हैं जो R में नहीं चलेंगे।



आम त्रुटि संदेश आर में

यदि आपको लाल आउटपुट मिलता है, तो आपने एक त्रुटि का अनुभव किया है। यहां कुछ सबसे आम त्रुटि संदेश दिए गए हैं जिनका आप सामना करेंगे।

Error: Object '...' not found

इसका मतलब है कि संदर्भित वस्तु आपके कार्यक्षेत्र में नहीं है। यह हो सकता है क्योंकि:

1. आप डेटा लोड करना या पैकेज स्थापित करना भूल गए।

2. आपने टाइप-ओ या कैपिटलाइज़ेशन त्रुटि की है।

3. आप उद्धरण चिह्नों को भूल गए होंगे, उदा। परिकल्पना परीक्षणों के लिए फ़ंक्शन इनपुट के रूप में "अधिक"। यह भी जाँचें कि तार्किक जैसे TRUE / FALSE सभी कैप में हैं।

4. आप पहले कोड की एक पंक्ति चलाना भूल गए थे, जिस ऑब्जेक्ट का आप उल्लेख कर रहे हैं। (सुनिश्चित करने के लिए अपने कंसोल से स्क्रॉल करें)।

5. आप एक विशिष्ट डेटासेट के भीतर एक चर को संदर्भित करने का प्रयास कर सकते हैं।

आपने एक प्लॉट बनाया है, लेकिन यह आपके प्लॉट विंडो में फिट नहीं है। प्लॉट विंडो का आकार बढ़ाने की कोशिश करें और अपने प्लॉट कमांड को फिर से रन करें।

Error: unexpected numeric constant in: ...

आपको सबसे अधिक संभावना है कि एक कोष्ठक, एक अल्पविराम याद आ रहा है, या जब आप पिछली पंक्ति को पूरा कर चुके थे, तो आपको संकेत के साथ कोड की एक पंक्ति भागा। अपनी कोड लाइन को ध्यान से पढ़ें और सभी उचित सिंटैक्स की जाँच करें।

Error in ....: undefined columns selected

इसका अर्थ है कि आपके द्वारा चयनित डेटा का कॉलम मौजूद नहीं है। यदि संख्यात्मक रूप से कॉलम का चयन करना है, तो सुनिश्चित करें कि आपके पास सूचकांक सही हैं। यदि नाम से चयन किया जाता है, तो वर्तनी और पूंजीकरण की जांच करें। अंत में, यह सुनिश्चित करने के लिए जांचें कि आपने डेटा को सही ढंग से लोड किया है और यह कि चर नाम कॉलम कॉलम के रूप में दिखाई दे रहे हैं और डेटा की पहली पंक्ति के रूप में नहीं।

संदर्भ

R Core Team (2012). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/

RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

R सॉफ़्टवेयर का उपयोग कर सांख्यिकीय तकनीक

उपेंद्र कुमार प्रधान और मलिक फरमान खान
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
upendra.pradhan@icar.gov.in

1. परिचय: (Introduction)

R सिर्फ एक प्रोग्राम से ज्यादा है जो आंकड़े करता है। यह सांख्यिकीय कंप्यूटिंग और ग्राफिक्स के लिए एक परिष्कृत कंप्यूटर भाषा और वातावरण है। R सांख्यिकीय कंप्यूटिंग वेबसाइट (www.r-project.org) के लिए आर-प्रोजेक्ट से उपलब्ध है। R एक खुला स्रोत (जीपीएल) सांख्यिकीय वातावरण है जिसे एस और एस-प्लस के बाद तैयार किया गया है S भाषा को 1980 के दशक के अंत में AT. में विकसित किया गया था आर परियोजना 1995 में ऑकलैंड विश्वविद्यालय के सांख्यिकी विभाग के रॉबर्ट जेंटलमैन और रॉस इहाका (इसलिए नाम, आर) द्वारा शुरू की गई थी इसने तेजी से व्यापक दर्शक वर्ग प्राप्त किया है। यह वर्तमान में आर कोर-डेवलपमेंट टीम द्वारा बनाए रखा जाता है, जो स्वयंसेवी डेवलपर्स की एक मेहनती, अंतरराष्ट्रीय टीम है। आर प्रोजेक्ट वेबपेज आर के बारे में जानकारी के लिए मुख्य साइट है। इस साइट पर सॉफ्टवेयर, साथ में पैकेज, और दस्तावेज़ीकरण के अन्य स्रोतों को प्राप्त करने के निर्देश हैं। R एक शक्तिशाली सांख्यिकीय कार्यक्रम है लेकिन यह सबसे पहले और सबसे महत्वपूर्ण प्रोग्रामिंग भाषा है। दुनिया भर के लोगों द्वारा R के लिए कई रूटीन लिखे गए हैं और R प्रोजेक्ट वेबसाइट से "पैकेज" के रूप में स्वतंत्र रूप से उपलब्ध कराए गए हैं। हालाँकि, मूल स्थापना (लिनक्स, विंडोज या मैक के लिए) में अधिकांश उद्देश्यों के लिए उपकरणों का एक शक्तिशाली सेट होता है। क्योंकि R एक कंप्यूटर भाषा है, यह उन अधिकांश प्रोग्रामों से थोड़ा भिन्न रूप से कार्य करता है जिनसे उपयोगकर्ता परिचित हैं। आपको कमांड टाइप करना होगा, जिनका मूल्यांकन प्रोग्राम द्वारा किया जाता है और फिर निष्पादित किया जाता है। यह कई उपयोगकर्ताओं के लिए थोड़ा कठिन लगता है, लेकिन R भाषा को चुनना आसान है और बहुत सारी सहायता उपलब्ध है। अन्य अनुप्रयोगों से कमांड में कॉपी और पेस्ट करना संभव है (उदाहरण के लिए: वर्ड प्रोसेसर, स्प्रेडशीट, या वेब ब्राउज़र) और यह सुविधा बहुत उपयोगी है, खासकर यदि आप सीखते समय नोट्स रखते हैं। इसके अतिरिक्त, R के विंडोज और मैकिंटोश संस्करणों में एक ग्राफिकल यूजर इंटरफेस (जीयूआई) है जो कुछ बुनियादी कार्यों में मदद कर सकता है।
R जटिल सांख्यिकीय उपागमों को उतनी ही सरलता से संभालता है जितना कि सरल। इसलिए, एक बार जब आप R भाषा की मूल बातें जान लेते हैं, तो आप जटिल विश्लेषणों को उतनी ही सरलता से हल कर सकते हैं (हमेशा की तरह यह परिणामों की व्याख्या है जो वास्तव में कठिन बिट हो सकता है)। अगले भाग में हम R का प्रयोग करते हुए कुछ बुनियादी सांख्यिकीय तकनीकों पर चर्चा करेंगे।

2. सारांश आँकड़े: (Summary statistics)

सारांश आँकड़े हमें टिप्पणियों (डेटा) के एक समूह को संक्षेप में प्रस्तुत करने में मदद करते हैं, ताकि कुछ संख्याओं में अधिक से अधिक जानकारी संप्रेषित की जा सके। एक आँकड़ा एक नमूना संपत्ति है, अर्थात, इसकी गणना नमूने में टिप्पणियों से की जा सकती है। इसके विपरीत, एक पैरामीटर जनसंख्या की एक संपत्ति है जिससे नमूना लिया गया था। चूंकि यह आमतौर पर अज्ञात होता है (जब तक कि हम किसी ज्ञात वितरण से डेटा का अनुकरण नहीं करते),

हम मापदंडों का अनुमान लगाने के लिए आंकड़ों का उपयोग करते हैं। आंकड़े हमें नमूने में देखे गए मूल्यों के वितरण के बारे में सूचित करते हैं। R में बहुत सारे आँकड़ों की गणना आसानी से की जा सकती है। यहाँ कुछ आँकड़ों का अवलोकन दिया गया है।

Statistic	R-function	Formula	Parameter
arithmetic mean	mean()	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$	population mean (μ)
median	median()	$median = x_{(n+1)/2}\ (for\ uneven\ n)$ $median = \frac{1}{2}\left(x_{n/2} + x_{(n/2)+1}\right)\ (for\ even\ n)$	population median
sample variance	var()	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\bar{x} - x_i)^2$	population variance (σ^2)
sample standard deviation	sd()	$s = \sqrt{s^2}$	population standard deviation (σ)

2.1. केंद्रीय प्रवृत्ति के उपाय: *(Measures of central tendency)*

केंद्रीय प्रवृत्ति का सबसे महत्वपूर्ण माप अंकगणितीय माध्य (या औसत) है। सममित वितरण (जैसे सामान्य वितरण) के "केंद्र" का वर्णन करना बहुत उपयोगी है। इसका नुकसान यह है कि यह चरम मूल्यों के प्रति संवेदनशील है। माध्यिका स्थान का एक वैकल्पिक उपाय है जो चरम मूल्यों के प्रति बहुत कम संवेदनशील होता है। यह आदेशित नमूने का केंद्रीय मूल्य है (तालिका में दिया गया सूत्र केवल तभी लागू होता है जब नमूना का आदेश दिया जाता है)। यदि n सम है, तो यह दो सबसे केंद्रीय मूल्यों का अंकगणितीय माध्य है।
आइए R में कुछ डेटा का अनुकरण करें और माध्य और माध्यिका दोनों की गणना करें:

```
x <- rnorm(20, mean = 5, sd = 4)          # x: sample data, n=20
x2 <- c(x, 45)                # add an extreme value of 45 to x     mean(x);
           mean(x2)
```

आप देखते हैं कि x का अंकगणितीय माध्य 5 के करीब है, सामान्य वितरण का सही मतलब हमने 20 मानों का नमूना लिया है। x2 का माध्य, नमूना x एक चरम मान सहित काफी बढ़ गया है। हालांकि, औसत चरम मूल्य से केवल थोड़ा ही बदलता है।

```
Median(x)

median(x2)

z <- c(1,4,7,9,10)   # small sample with uneven n
median(z)
```

```
z2 <- c(1,4,7,9)  # small sample with even n
median(z2)
```

एक पूर्ण सममित वितरण के लिए, माध्यिका माध्य के बराबर होती है।

2.2. फैलाव के उपाय: *(Measures of dispersion)*

फैलाव के उपाय डेटा के प्रसार / परिवर्तनशीलता को मापते हैं। एक नमूने का प्रसरण नमूना में सभी टिप्पणियों पर नमूना माध्य से वर्ग विचलन का योग है, जिसे (n-1) से विभाजित किया जाता है। विचरण की व्याख्या करना कठिन है, क्योंकि यह आमतौर पर काफी बड़ी संख्या होती है। मानक विचलन, जो विचरण का वर्गमूल है, आसान है।
यह लगभग नमूना माध्य से एक अवलोकन का औसत विचलन है (यह नमूना माध्य से बिल्कुल औसत विचलन होगा, यदि हम मानक विचलन के लिए सूत्र के हर में n-1 के बजाय n का उपयोग करेंगे)। ऊपर बनाए गए नमूना x के लिए प्रयास करें:

```
var(x)            # sample variance
sd(x)             # sample standard deviation
sqrt(var(x))      # dito
```

2.3. क्वांटाइल्स और बॉक्सप्लॉट: *(Quantiles and the boxplot)*

p-क्वांटाइल वह मान x है, जिसके गुण से कम या उसके बराबर मान प्राप्त होने की प्रायिकता p है, अर्थात p(X≤ x) = p. माध्यिका 50% मात्रा है। 25% क्वांटाइल और 75% क्वांटाइल को निचला और ऊपरी क्वार्टाइल भी कहा जाता है (चतुर्थक क्योंकि माध्यिका के साथ मिलकर, वे वितरण को क्वार्टरों में विभाजित करते हैं)। 25% और 75% चतुर्थक के बीच के अंतर को अंतर-चतुर्थक श्रेणी कहा जाता है। इस श्रेणी में वितरण का 50% शामिल है और इसका उपयोग फैलाव के माप के रूप में भी किया जाता है। हम पहले ही देख चुके हैं कि द्विपद, पॉइसन और सामान्य वितरण (फ़ंक्शन qbinom (), qpois (), qnorm ()) के लिए चतुर्थक की गणना कैसे की जाती है। . मानक सामान्य वितरण के 25%, 50% और 75% मात्रा की गणना निम्नानुसार की जा सकती है:

```
q25 <- qnorm(0.25); q25
q50 <- qnorm(0.5); q50
q75 <- qnorm(0.75); q75
```

बॉक्सप्लॉट माध्यिका, चतुर्थक और अंतर-चतुर्थक श्रेणी का उपयोग करके वितरण को रेखांकन रूप से प्रदर्शित करने का एक तरीका है। आइए पुस्तकालय ISwR से डेटासेट tlc लोड करें और महिलाओं और पुरुषों के शरीर के आकार का एक बॉक्सप्लॉट बनाएं।

```
library(ISwR)
data(tlc);      ?tlc
par(mfrow=c(1,1))

boxplot(height ~ sex, data = tlc, names=c("Women","Men"), ylab = "Body
height (cm)", col = "blue")
```

एक बॉक्सप्लॉट में, बॉक्स (यहाँ नीले रंग में) में इंटरक्वेर्टाइल रेंज शामिल है। माध्यिका को बॉक्स के भीतर एक रेखा के रूप में दिखाया गया है। व्हिस्कर्स द्वारा परिभाषित रेंज जो बॉक्स के ऊपरी और निचले सिरे पर फैली हुई है, कम स्पष्ट रूप से परिभाषित (सॉफ्टवेयर और उपयोगकर्ता पर निर्भर) है। R में डिफ़ॉल्ट (?boxplot.stats देखें) यह है कि मूंछें सबसे चरम डेटा बिंदु तक फैली हुई हैं जो बॉक्स से दूर बॉक्स की लंबाई से 1.5 गुना से अधिक नहीं है (इंटरक्वेर्टाइल-रेंज)। इस सीमा से परे डेटा बिंदु अलग से "आउटलेयर" के रूप में दिखाए जाते हैं (नीचे चित्र देखें)।

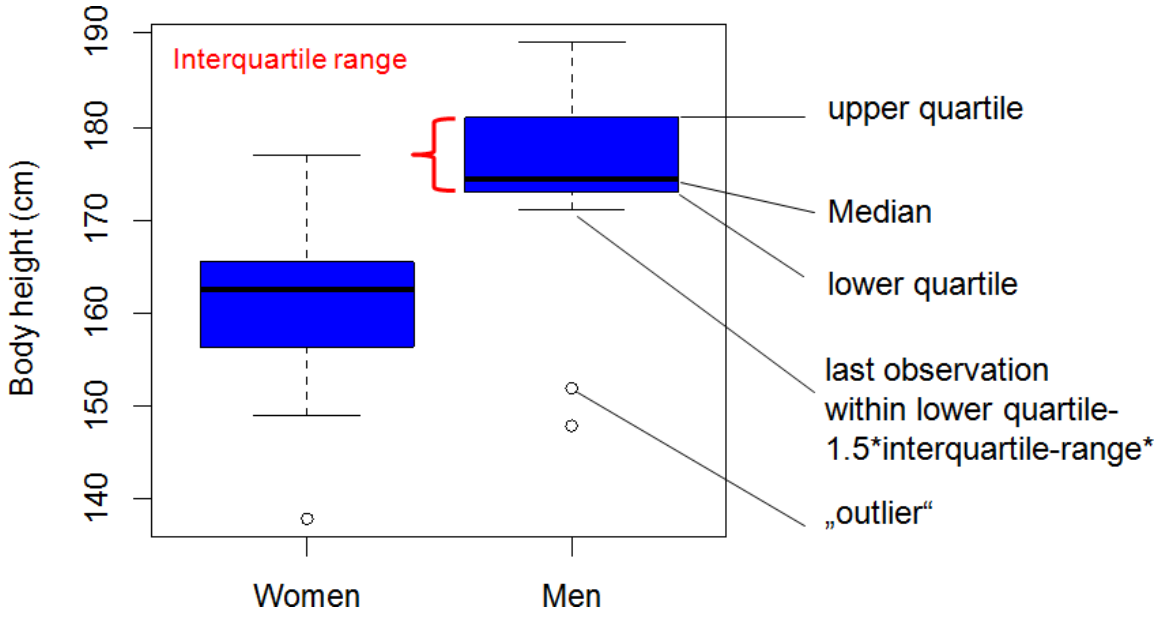


Figure: Comparison of body height of women and men by box plots.

1. 2.4. माध्य की मानक त्रुटि: (The standard error of the mean)

यदि हमारे पास n प्रेक्षणों का एक नमूना है, जो सभी माध्य μ के साथ एक सामान्य वितरण से आते हैं और मानक विचलन σ,तो यह ज्ञात है कि नमूने का अंकगणितीय माध्य $\bar{x}$ सामान्य रूप से μ के आसपास मानक विचलन $SD_{\bar{x}} = \sigma/\sqrt{n}$. के साथ वितरित किया जाता है।

व्यवहार में, हालांकि, हम नहीं जानते हैं, लेकिन नमूना मानक विचलन s द्वारा इसका अनुमान लगाते हैं। नमूना माध्य μ के वास्तविक मानक विचलन का अनुमान "माध्य की मानक त्रुटि" (SEM) द्वारा भी लगाया जाता है, जिसकी गणना $SEM = \sigma/\sqrt{n}$. के रूप में की जाती है। यहाँ एक उदाहरण है।

```
x <- rnorm(n = 100, s = 4, mean = 20)      # a sample of size n=100
sem <- sd(x) / sqrt(length(x)); sem        # calculate SEM
sd.mean <- 4/sqrt(100)                     # theoretical SD.mean
```

जबकि s व्यक्तिगत टिप्पणियों की परिवर्तनशीलता का वर्णन करता है (नमूना माध्य से औसत अंतर, SEM नमूना मानक विचलन s के साथ नमूने से n यादृच्छिक मानों के नमूना माध्य की परिवर्तनशीलता का वर्णन करता है।

2.5. विश्वास अंतराल: *(Confidence intervals)*

यदि (ऊपर के रूप में) हमारे पास n अवलोकनों का एक नमूना है, जो सभी माध्य μ के साथ सामान्य वितरण से आते हैं और मानक विचलन , σ और हम जानते हैं कि अंकगणितीय माध्य $\bar{x}$ नमूना का सामान्य रूप से आसपास μ वितरित किया जाता है मानक विचलन के साथ $SD_{\bar{x}} = \sigma/\sqrt{n}$ हम एक 95% की गणना कर सकते हैं विश्वास अंतराल के लिए μ जैसा $\bar{x} + \sigma/\sqrt{n}.N_{0.025} \leq \mu \leq \bar{x} + \sigma/\sqrt{n}N_{0.975}$, जहां $N_{0.025}$ और $N_{0.975}$ मानक सामान्य वितरण के 2.5% क्वांटाइल हैं। उपरोक्त उदाहरण के लिए यह होगा:

```
x <- rnorm(n = 100, s = 4, mean = 20)    # a sample of size n=100
sample.mean <- mean(x)
sd.mean <- 4/sqrt(100)                           # theoretical SD.mean
ci95.lower <- sample.mean + sd.mean * qnorm(0.025)
ci95.upper <- sample.mean + sd.mean * qnorm(0.975)
```

95% विश्वास अंतराल के लिए μ एक प्रशंसनीय सीमा निर्दिष्ट करता है। यदि हम सच्चे पैरामीटर के लिए एक विश्वास अंतराल का अनुमान लगाते हैं μ अनंत संख्या में, नमूनों की अनंत संख्या μ से की गणना करके, प्रत्येक आकार ν, तो 95% बार (जैसे, 1000 में से लगभग 50 बार) μ विश्वास अंतराल के भीतर होगा। 95% विश्वास अंतराल का अर्थ कभी-कभी 95% संभावना के साथ μ युक्त होने के रूप में गलत होता है। यह मामला नहीं है, क्योंकि एक विशिष्ट विश्वास अंतराल में μ या तो होता है या यह नहीं होता है। हालांकि, व्यवहार में हम नहीं जानते हैं और नमूने से σ का अनुमान लगाना पड़ता है। सामान्य वितरण का उपयोग करने के बाद एक आत्मविश्वास अंतराल होता है जो बहुत संकीर्ण होता है। इसलिए डेटा से अनुमानित विश्वास अंतराल सामान्य वितरण के बजाय टी-वितरण (नीचे

देखें) पर आधारित होते हैं। नमूने से σ अनुमानित होने के लिए इसे सही करना आवश्यक है, खासकर यदि नमूना आकार छोटा है।

3. शास्त्रीय सांख्यिकीय परीक्षण: (Classical statistical tests)

3.1. शून्य-परिकल्पना परीक्षण: *(Null-hypothesis testing)*

शास्त्रीय सांख्यिकीय विधियां अशक्त-परिकल्पना परीक्षण के साथ काम करती हैं जो मिथ्याकरण पर आधारित है। तार्किक दृष्टिकोण से, एक सिद्धांत या एक व्युत्पन्न परिकल्पना को साबित करना असंभव है, क्योंकि इसके लिए परिकल्पना से संबंधित सभी टिप्पणियों की आवश्यकता होगी, और हम आमतौर पर नमूनों के साथ काम करते हैं। इस प्रकार मिथ्याकरण, परिकल्पनाओं को सिद्ध करने के बजाय, पूछताछ के लिए, मिथ्याकरण के लिए प्रयास करता है।

लंबे समय तक, उदाहरण के लिए, यूरोप में केवल सफेद हंस देखे गए थे और यह अनुमान लगाया गया था कि सभी हंस सफेद होते हैं। हालाँकि, ऑस्ट्रेलिया में एक एकल काले हंस का अवलोकन इस परिकल्पना का खंडन करने के लिए पर्याप्त था। शास्त्रीय सांख्यिकीय परीक्षण करते समय, हम एक चक्कर लगाते हैं। हम एक अशक्त-परिकल्पना (H_0) का प्रस्ताव करते हैं जो आम तौर पर रुचिकर नहीं होती है और हमारी वास्तविक परिकल्पना को साबित करने के बजाय इसे गलत साबित करने (अस्वीकार) करने का प्रयास करते हैं। H_0एक वैकल्पिक परिकल्पना (H_A) द्वारा पूरित है, जो उदाहरण के लिए, हमने उत्तरी फुलमार (ईस्टुरमवोगेल, फुलमारस ग्लेशियलिस) की चयापचय दर को मापा होगा और यह जानना चाहते हैं कि क्या पुरुष और महिलाएं चयापचय दर में भिन्न हैं। संगत H_0यह होगा कि पुरुषों और महिलाओं की चयापचय दर समान होती है। हा यह होगा कि चयापचय दर भिन्न होती है। ध्यान दें कि इस मामले में H_A दो तरफा है, जिसका अर्थ है कि हम यह निर्दिष्ट नहीं करते हैं कि पुरुषों या महिलाओं की चयापचय दर अधिक है या नहीं। H_0 का परीक्षण करने के लिए, हम प्रेक्षित होने की प्रायिकता का अनुमान लगाते हैं यदि P-मान $< \mu$ है (नीचे देखें), तो हम H_0 को अस्वीकार करते हैं। हम कहते हैं, परीक्षण महत्वपूर्ण है।

		Truth	
		H_0 is true	**H_A is true**
Decision (test result)	**accept H_0**	1-α	type II error (β)
	reject H_0	type I error (α)	1-β

3.1.1. टेस्ट आँकड़े: *(Test statistics)*

जैसा कि पहले ही कहा जा चुका है, संभावना है कि देखे गए परिणाम या अधिक चरम परिणाम अकेले संयोग से हो सकते हैं (यदि एच 0 सत्य थे), पी-मान है। पी-मान आमतौर पर एक परीक्षण-सांख्यिकी और उस परीक्षण आंकड़े के वितरण के माध्यम से प्राप्त होता है। एक परीक्षण आँकड़ा का एक उदाहरण t परीक्षण के मामले में t आँकड़ा और t-

वितरण है उदाहरण के लिए अन्य परीक्षण आँकड़े हैं। $\chi 2$ ($\chi 2$ परीक्षण, संभावना-अनुपात परीक्षण या F (विचरण के विश्लेषण में एफ-परीक्षण)।

3.2. टी टेस्ट परिवार: *(The t test family)*

तीन प्रकार के टी परीक्षण हैं जिन पर हम गौर करने जा रहे हैं, जिनका उपयोग हम जिस तरह की तुलना करना चाहते हैं, उसके आधार पर किया जाता है। ये सभी इस धारणा पर आधारित हैं कि नमूना डेटा सामान्य वितरण से आता है। इसका मतलब यह भी है कि टी परीक्षण निरंतर डेटा के लिए हैं।

आइए एक-नमूना टी परीक्षण से शुरू करें।

3.2.1. एक-नमूना टी परीक्षण: *(One-sample t test)*

एक-नमूना t परीक्षण का उपयोग एक माध्य (नमूने के) की तुलना संदर्भ मान (एक प्राथमिकता से चुना गया मान) से करने के लिए किया जाता है। डालगार्ड (2008) से एक उदाहरण यहां दिया गया है: हमारे पास: 11 महिलाओं से दैनिक ऊर्जा सेवन (जूल में) के माप का एक नमूना, $x_1,x_2,...,x_{11}$। अब हम जानना चाहते हैं कि क्या ये डेटा प्रति दिन 7725 जूल के अनुशंसित मूल्य के अनुरूप हैं?

```
daily.intake <- c(5260,5470,5640,6180,6390,6515,6805,7515,7515,8230,8770)
Mean <- mean(daily.intake); Mean
```

शून्य परिकल्पना, H_0, है: $\mu = \mu_0$, वैकल्पिक परिकल्पना, H_A: $\mu \neq \mu_0$ इस प्रकार, वास्तविक μ जनसंख्या माध्य है जिसका अनुमान हम प्रतिदर्श माध्य $\bar{x}$ से लगाते हैं। μ_0 संदर्भ मान है (यहाँ प्रशंसित ऊर्जा सेवन)। यदि यह धारणा कि हमारा नमूना सामान्य रूप से माध्य μ और विचरण σ^2 अर्थात $x_1,x_2,...,x_n \sim N(\mu, \sigma^2)$, के साथ वितरित किया जाता है, उचित है, तो हम अपने प्रश्न का उत्तर देने के लिए एक-नमूना t परीक्षण का उपयोग कर सकते हैं। . इस परीक्षण के लिए एक प्रमुख अवधारणा माध्य की मानक त्रुटि है $SEM = s/\sqrt{n}$ जो हम पिछले अध्याय में मिल चुके हैं। t परीक्षण का परीक्षण आँकड़ा t है (क्या आश्चर्य है), और इसकी गणना इस प्रकार की जाती है:

$$\frac{\bar{x} - \mu_0}{SEM}$$

t मूल रूप से मापता है कि हमारे नमूना माध्य कितने SEM संदर्भ मान से भिन्न हैं। याद रखें कि सामान्य रूप से वितरित डेटा के लिए, $\mu \pm 2\sigma$ के भीतर रहने की संभावना लगभग 95% है (इस सीमा के बाहर 5% के अनुरूप)। H_0 के तहत, हम इस सीमा के भीतर μ_0 के गिरने की उम्मीद करेंगे। छोटे नमूनों में (मान लीजिए n <30), हालांकि, इस तथ्य के लिए सही करना आवश्यक है कि हमने नमूने से SEM का अनुमान लगाया है। इसलिए परीक्षण सांख्यिकीय t (t-वितरण) के वितरण में चरम मूल्यों के लिए संभावनाओं में थोड़ी वृद्धि हुई है, अर्थात, t-वितरण में भारी पूंछ होती है। t-वितरण का आकार नमूना आकार पर निर्भर करता है और n से अनंत तक बढ़ने के लिए सामान्य वितरण का अनुमान लगाता है (नीचे चित्र देखें)। H_0 के स्वीकृति क्षेत्र को परिभाषित करने के लिए सही मात्रा प्राप्त करने के लिए हम उन्हें n-1 डिग्री स्वतंत्रता के साथ t- वितरण के लिए लेते हैं। यदि हमारा मनाया गया टी किसी दिए गए महत्व स्तर के लिए स्वीकृति क्षेत्र से बाहर आता है, यानी $[t_{0.025}, n\text{-}1/t_{0.975}, n\text{-}1]$ for α=5 %, के लिए, हम H_0 को अस्वीकार करते हैं और कहते हैं कि नमूना माध्य काफी भिन्न होता है H_0 से। समान रूप से, हमें इस परीक्षण के लिए एक p-मान <0.05 प्राप्त होगा, जो कि t-मान को देखे गए मान से बड़े या बड़े के रूप में देखने की संभावना है, दिया गया H_0 सत्य है। ध्यान दें कि बड़े नमूनों ($n \geq 30$) के लिए, आप अंगूठे के नियम का उपयोग कर सकते हैं: $t \geq 2 \rightarrow \mu \neq \mu_0$ अब हम केवल हाथ से वर्णित गणना करते हैं:

```
# One sample t test by hand
Mu0 <- 7725

SD <- sd(daily.intake); SD

[1] 1142.123

N <- length(daily.intake); N

[1] 11

SEM <-SD/sqrt(N); SEM                    # standard error of the mean

[1] 344.3631

Tval <- (Mean-Mu0)/SEM; Tval             # to calculate t manually

[1] -2.820754

pvalue <- 2*pt(Tval, N-1); pvalue

[1] 0.01813724
```

R में t.test() फ़ंक्शन का उपयोग करके यह परीक्षण करना बहुत तेज़ है:

```
  # One-sample t test using t.test()

  > t.test(daily.intake, mu = 7725)# 7725: recommended valueOne Sample t-test

data:  daily.intake

t = -2.8208, df = 10, p-value = 0.01814

alternative hypothesis: true mean is not equal to 7725

95 percent confidence interval:

   5986.348 7520.925

sample estimates:

mean of x
6753.636
```

हमें वही परिणाम मिलता है जो हमें हाथ से मिला था, लेकिन अधिक सुसंगत आउटपुट के साथ। हमें उपयोग किया गया डेटा सेट, मनाया गया टी-मान, स्वतंत्रता की डिग्री और पी-मान, टी-वितरण से देखे गए टी-मान (या अधिक चरम एक) की संभावना प्राप्त होती है। इसके अलावा, R हमें हमारे H_A के बारे में याद दिलाता है और सूत्र "बराबर नहीं" हमें इस तथ्य की याद दिलाता है कि हमारे पास दो-तरफा H_A है। R के लिए μ 95% विश्वास अंतराल भी प्रदान करता है जिसे हम हाथ से इस प्रकार गणना कर सकते हैं:

$$\bar{x} + t_{0.025,\ n-1} \times SEM < \mu < \bar{x} + t_{0.975,\ n-1} \times SEM$$

3.2.2 दो-नमूना टी परीक्षण: *(The two-sample t test)*

एक माध्य को संदर्भ मान से तुलना करने के बजाय, हम अक्सर दो साधनों (दो नमूने) की तुलना करना चाहते हैं। हम यह भी कह सकते हैं कि हम नल-परिकल्पना का परीक्षण करना चाहते हैं कि दो नमूने समान माध्य के साथ वितरण से आते हैं। यहाँ Quinn . से एक उदाहरण है उन्होंने उत्तरी फुलमार (ईस्टुरमवोगेल, फुलमारस ग्लेशियलिस) की चयापचय दर को मापा है और यह जानना चाहते हैं कि क्या पुरुष और महिलाएं चयापचय दर में भिन्न हैं। संबंधित शून्य परिकल्पना, H_0, है: $\mu_M = \mu_F$ या समकक्ष $\mu_M - \mu_F = 0$, और (दो तरफा) वैकल्पिक परिकल्पना, H_A है: $\mu_M - = \mu_F$ या: $\mu_M - \mu_F \neq 0$.हम मानते हैं कि पुरुष और महिला दोनों चयापचय दर सामान्य रूप से वितरित की जाती हैं, अर्थात, पुरुष चयापचय दर ~ N (μM, μM^2) और महिला चयापचय दर ~ N (μF, μF^2)। एक और महत्वपूर्ण धारणा यह है कि दो नमूने स्वतंत्र हैं इसका मतलब है कि चयापचय दर का प्रत्येक माप एक स्वतंत्र "इकाई" (यहाँ पक्षी) पर लिया गया था। हम अगले उप-अध्याय में गैर-स्वतंत्र नमूनों का एक उदाहरण देखेंगे। दो-नमूना टी परीक्षण उसी तरह से काम करता है जैसे कि एक-नमूना t परीक्षण। परीक्षण आँकड़ा t है:

$$\frac{\bar{x}_2 - \bar{x}_1}{SEDM}$$

SEDM साधनों के अंतर की मानक त्रुटि है। इस धारणा के तहत कि σ_M और σ_F बराबर हैं, इसकी गणना माध्य (शास्त्रीय टी परीक्षण) की जमा मानक त्रुटि के रूप में की जाती है:

$$SEDM = s_p \sqrt{\frac{1}{n1} + \frac{1}{n2}}$$

$$sp = \sqrt{(n1 - 1)\, s2\,1 + (n2 - 1)\, s2\,2\,]/\, n1 + n2 - 2}$$

संबंधित टी स्वतंत्रता के n1 +n2-2 डिग्री के साथ एक टी-वितरण का पालन करेगा।
हालांकि, समान भिन्नताओं की धारणा हमेशा उपयुक्त नहीं होती है, और R डिफ़ॉल्ट रूप से वेल्च परीक्षण प्रदान करता है, जो यह धारणा नहीं बनाता है। परिणामी t का वितरण s1, s2, n1, और n2 (गैर-पूर्णांक df) से परिकलित स्वतंत्रता की डिग्री के साथ t वितरण द्वारा अनुमानित किया जा सकता है। वेल्च परीक्षण को सबसे सुरक्षित माना जाता है। लेकिन जब तक समूह के आकार और मानक विचलन बहुत भिन्न नहीं होते हैं, तब तक दो परीक्षण आमतौर पर समान परिणाम देंगे। अब आइए R में दो-नमूना टी परीक्षण के इन प्रकारों को देखें। सबसे पहले, आपको डेटा सेट "furness.csv" को पढ़ना होगा, उदाहरण के लिए अपने फ़ोल्डर "डेटासेट" में कार्यशील निर्देशिका सेट करके:

```
setwd("..your filepath…/datasets")# set the working directory

furness <- read.csv("furness.csv")          # read the data
```

किसी भी परीक्षण को करने से पहले डेटा का ग्राफिक रूप से निरीक्षण करना हमेशा एक अच्छा विचार है। फ़र्नेस डेटा के लिए सेक्स द्वारा एक बॉक्सप्लॉट ऐसा करने का एक अच्छा तरीका है:

```
  plot(METRATE ~ SEX, data = furness)
```

ग्राफिक से पता चलता है कि प्रसरण असमान हैं। हम डिफ़ॉल्ट वेल्च परीक्षण से शुरू करते हैं:

```
  t.test(METRATE ~ SEX, data = furness)     # default: Welch two-sample t test
```

```
Welch Two Sample t-test

data:  METRATE by SEX

t = -0.7732, df = 10.468, p-value = 0.4565

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1075.3208   518.8042

sample estimates:

mean in group Female   mean in group Male
            1285.517   1563.775
```

शास्त्रीय दो-नमूना टी परीक्षण प्राप्त करने के लिए हमें स्पष्ट रूप से कहना होगा कि भिन्नताएं समान मानी जाती हैं:

```
t.test(METRATE ~ SEX, data = furness, var.equal = T) # classical t test,
assuming equal variances

Two Sample t-test

data:  METRATE by SEX

t = -0.7009, df = 12, p-value = 0.4968

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1143.3057   586.7891

sample estimates:

mean in group Female   mean in group Male
            1285.517   1563.775
```

दोनों टेस्ट वेरिएंट का आउटपुट एक-नमूना टी टेस्ट के समान है। लेकिन विश्वास अंतराल अब साधनों में अंतर के लिए है। शून्य, H_0 के तहत अंतर, 95% आत्मविश्वास अंतराल में निहित है, यह सुझाव देता है कि नर और मादा फुलमार चयापचय दर में भिन्न (ज्यादा) नहीं होते हैं। वही p-मान> 0.05 द्वारा सुझाया गया है। शास्त्रीय परीक्षण के लिए स्वतंत्रता की डिग्री $n_F + n_M - 2 = 6 + 8 – 2 = 12$. हैं। ध्यान दें कि विनिर्देश METRATE ~ SEX एक मॉडल सूत्र (सेक्स द्वारा समझाया गया चयापचय दर) से मेल खाता है और आमतौर पर डेटा फ़्रेम में संग्रहीत डेटा के उपयोग के लिए बहुत आसान है। वैकल्पिक रूप से, हम दोनों समूहों को अलग-अलग निर्दिष्ट कर सकते हैं, लेकिन यह लिखना अधिक जटिल है:

```
# alternative specification
t.test(furness$METRATE[furness$SEX == "Female"],

       furness$METRATE[furness$SEX == "Male"], var.equal = T)
```

हम आपको ग्राफिक्स का उपयोग करने की सलाह देते हैं (नीचे दिए गए बॉक्सप्लॉट के रूप में) और यह तय करने के लिए कि आप किस परीक्षण का उपयोग करना चाहते हैं, आपके अपने तर्क हैं। R, var.test() में भिन्नताओं की समानता के लिए एक औपचारिक परीक्षण भी है, जिसे हम यहां जल्दी से आजमाते हैं:

```
var.test(METRATE ~ SEX, data = furness)
```

गैर-महत्वपूर्ण परीक्षा परिणाम बताता है कि भिन्नताएं समान हो सकती हैं (अलग नहीं)। हालांकि, यह परीक्षण किसी भी सांख्यिकीय परीक्षण के समान ही दोष से ग्रस्त है: छोटे नमूनों के लिए, आपको एक गैर-महत्वपूर्ण परिणाम प्राप्त होने की संभावना है (सांख्यिकीय शक्ति की कमी के कारण), इस मामले में आप समान भिन्नताओं की धारणा को अस्वीकार करने के लिए प्रेरित करते हैं। यदि अंतर वास्तव में काफी बड़ा है (जैसा कि हम बॉक्स प्लॉट से निष्कर्ष निकाल सकते हैं)। बड़े नमूनों के लिए, आपको एक महत्वपूर्ण परीक्षा परिणाम प्राप्त करने और समान भिन्नताओं (उच्च शक्ति) की धारणा को अस्वीकार करने की बहुत संभावना है। इस कारण से हम भिन्नताओं को आलेखीय रूप से आंकना पसंद करते हैं।

3.2.3. युग्मित नमूनों के लिए टी परीक्षण: *(The t test for paired samples)*

युग्मित नमूने तब होते हैं जब एक ही प्रायोगिक इकाई पर दो माप होते हैं। इसका मतलब है कि ये दो माप स्वतंत्र नहीं हैं। उदाहरण के लिए, आपने 11 महिलाओं में दो बार, मासिक धर्म से पहले और बाद में दैनिक ऊर्जा सेवन को मापा होगा। एक बड़े मासिक धर्म मूल्य को एक बड़े मासिक धर्म मूल्य (उसी महिला के) के साथ जोड़े जाने की संभावना है। यह आकलन करने के लिए कि क्या मासिक धर्म से पहले और बाद की ऊर्जा का सेवन अलग-अलग है, हम प्रत्येक महिला के लिए दोनों के बीच के अंतर की गणना कर सकते हैं और शून्य के खिलाफ अंतर के नमूने की तुलना करते हुए एक-नमूना टी परीक्षण कर सकते हैं (बिना किसी अंतर के H_0 के तहत संदर्भ मान)। इस परीक्षण के लिए एक महत्वपूर्ण धारणा यह है कि मतभेदों का एक वितरण होता है जो स्तर से स्वतंत्र होता है (बड़े मूल्यों के बीच अंतर छोटे मूल्यों के बीच की तुलना में बड़ा अंतर नहीं होना चाहिए)। युग्मित डेटा के अन्य उदाहरण एक ही गमले में उगने वाले पौधों के जोड़े पर माप या भाई-बहनों के जोड़े पर माप हैं। स्वतंत्र और गैर-स्वतंत्र डेटा के बीच अंतर को समझना बहुत महत्वपूर्ण है। जब भी आपके पास "पदानुक्रमित" या "संकुल" डेटा होगा, तो आप उन्हें फिर से देखेंगे। आइए उदाहरण देखें और मासिक धर्म से पहले और बाद के मापों के बीच के अंतरों का रेखांकन करें। डेटा लाइब्रेरी ISwR से हैं जो Daalgard (2008) के साथ हैं:

```
library(ISwR)
data(intake)
# inspect differences graphically
difference <- intake$post - intake$pre
average <- (intake$post + intake$pre)/2
plot(average, difference)
```

सिर्फ n=11 महिलाओं का नमूना बहुत छोटा है। लेकिन ग्राफिक इंगित करता है कि ऊर्जा सेवन के स्तर से मतभेदों को स्वतंत्र मानना उचित है। और एक युग्मित t परीक्षण लागू करना सहेजा जाना चाहिए। । यह दोनों नमूनों का उपयोग करने और यह निर्दिष्ट करने के बराबर है कि वे युग्मित हैं, या हम केवल एक-नमूना टी-परीक्षण द्वारा अंतरों की तुलना शून्य से कर सकते हैं:

```
# using both samples and say they are paired

> t.test(intake$post, intake$pre , paired = TRUE)

        Paired t-test
```

```
data:  intake$post and intake$pre
t = -11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1566.838 -1074.072
sample estimates:
mean of the differences
              -1320.455
# alternative: use one-sample t test on difference
> t.test(difference, mu = 0)
One Sample t-test

data:  difference
t = -11.9414, df = 10, p-value = 3.059e-07
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1566.838 -1074.072
sample estimates:
mean of x
-1320.455
```

आउटपुट संख्यात्मक रूप से बराबर है, बस विवरण (डेटा, अंतर का माध्य बनाम x का माध्य) थोड़ा अलग है। ध्यान दें कि ऐसा युग्मित डिज़ाइन बहुत "कुशल" है यदि यह पूछे गए शोध प्रश्न के अनुकूल है (यानी, यदि आप उन चीजों में रुचि रखते हैं जो विषय के भीतर भिन्न हैं)। विषयों की जोड़ी के माध्यम से, आप मूल रूप से विषयों के बीच के अंतर से छुटकारा पा सकते हैं। डेटा की युग्मित प्रकृति को अनदेखा करने से इस मामले में सांख्यिकीय शक्ति कम हो जाएगी।. (हालांकि, अन्य सेटिंग्स में डेटा की जोड़ी या क्लस्टर प्रकृति को अनदेखा करने से विपरीत स्थिति हो सकती है, जिसे छद्म-प्रतिकृति भी कहा जाता है।

3.3. श्रेणीबद्ध डेटा के लिए परीक्षण: (Tests for categorical data)

3.3.1. किसी संदर्भ मान के अनुपात की तुलना करें: द्विपद परीक्षण: *(Compare a proportion to a reference value: the binomial test)*

आपने एक पेट्री डिश में 215 बीज डाले हैं और 39 अंकुरित हुए हैं। अब आप जांचना चाहते हैं कि क्या यह अंकुरण दर p = 39/215 व्यापारी द्वारा दी गई अंकुरण दर p0 = 0.15 के अनुरूप है। शून्य-परिकल्पना है: p = p0। आप इसे बिनोम.टेस्ट () फ़ंक्शन का उपयोग करके R में कर सकते हैं जो द्विपद वितरण बी (एन, पी) की संभावनाओं पर आधारित है:

```
binom.test(39,215,0.15)

        Exact binomial test

data:  39 and 215

number of successes = 39, number of trials = 215, p-value = 0.2135
alternative hypothesis: true probability of success is not equal to 0.15

95 percent confidence interval:

 0.1322842 0.2395223

sample estimates:

probability of success

          0.1813953
```

आउटपुट में अंकुरित होने की संभावना के लिए एक विश्वास अंतराल शामिल होता है। चूंकि कॉन्फिडेंस इंटरवल में p0 होता है, H_0 के तहत मान, प्रेक्षित अंकुरण दर व्यापारी द्वारा दी गई जानकारी के अनुरूप होती है। वैकल्पिक रूप से, आप इस परीक्षण को फ़ंक्शन prop.test() का उपयोग करके कर सकते हैं जो सामान्य सन्निकटन N(np,np(1-p)) का उपयोग करता है:

```
prop.test(39,215,0.15)

        1-sample proportions test with continuity correction data:  39 out of 215,

null probability 0.15
X-squared = 1.425, df = 1, p-value = 0.2326
alternative hypothesis: true p is not equal to 0.15

95 percent confidence interval:

 0.1335937 0.2408799

Sample estimates:

   p

0.1813953
```

परिणाम binom.test() के समान है क्योंकि n बड़ा है (n = 215)। डिफ़ॉल्ट रूप से, येट्स निरंतरता सुधार का उपयोग किया जाता है (उदाहरण के लिए अधिक विवरण के लिए डालगार्ड 2008 देखें)।

3.3.2. दो अनुपातों की तुलना करें: χ^2 परीक्षण (*Compare two proportions: χ^2 test*)

आपने पेट्री डिश में पौधों की प्रजाति A के 108 बीज और पौधे की प्रजाति B के 117 बीज बोए हैं। एक सप्ताह के बाद, प्रजाति A के 81 और प्रजाति B के 36 बीज अंकुरित हुए। क्या दो प्रजातियों के अंकुरण दर p_A और p_B. भिन्न हैं? संबंधित H_0 है: $p_A = p_B$. ऐसे कई परीक्षण हैं जिनका उपयोग इस उदाहरण के लिए किया जा सकता है। हम पहले χ^2 परीक्षण को देखेंगे और इसे R में "हाथ से" करेंगे। हम प्रेक्षित आवृत्तियों का एक मैट्रिक्स बनाकर शुरू करते हैं:

```
sown     <-     c(108,117)
germinated <- c(81,36)

notgerminated <- sown-germinated

observed <- matrix(c(notgerminated, germinated), ncol=2,
dimnames=list(c("species      A","species      B"),c("not
germinated","germinated"))); observed
```

```
            not germinated germinated

  species A       27           81
  species B       81           36
```

χ^2 परीक्षण मनाया आवृत्तियों की तुलना इस धारणा के तहत अपेक्षित आवृत्तियों के साथ करता है कि अंकुरण दर प्रजातियों से स्वतंत्र है (और इसके विपरीत, जो इस उदाहरण में थोड़ा जैविक अर्थ बनाता है)। अपेक्षित आवृत्तियों की गणना मैट्रिक्स की चार कोशिकाओं में से प्रत्येक के लिए सीमांत योग को गुणा करके की जाती है। प्रजाति A से अंकुरित बीजों के लिए यह होगा:

(all seeds of species A * all germinated seeds) / Total number of seeds (27+81) * (27 + 81) / (27 + 81 + 81 + 36) = 51.84

```
  # Calculate the marginal totals

  speciesA <- sum(observed["species A",])      # Total of species A
  speciesB <- sum(observed["species B",])      # Total of species B
  germ <- sum(observed[,"germinated"])         #   Total   germinated
  notgerm <- sum(observed[,"not germinated"]) # Total not germinated
  Total <- sum(observed)                       # Total seeds overall


  # calculate expected frequencies, assuming independence of species and
  germination rate

germA <- speciesA * germ / Total
germB <- speciesB * germ / Total

notgermA <- speciesA * notgerm / Total
notgermB <- speciesB * notgerm / Total#
matrix of expected frequencies

expected     <-     matrix(c(notgermA,     notgermB,     germA,     germB),
dimnames=list(c("species  A","species  B"),c("not  germinated","germinated")),
ncol = 2); expected
```

```
not germinated germinated

  species A          51.84       56.16

  species B          56.16       60.84


  calculate Chi-square observed

  Chi2 <- sum( ((observed - expected)^2) / expected ); Chi2
  [1] 44.01812

  # p-value

  pchisq(Chi2, df = 1,lower.tail = FALSE)

  [1] 3.253498e-11
```

परीक्षण के परिणाम पौधों की प्रजातियों और अंकुरण दर के बीच मजबूत निर्भरता को इंगित करते हैं। बेशक, R में एक फ़ंक्शन है जो गणना को बहुत तेजी से करता है। यदि हम इसका उपयोग करते हैं तो हमें हाथ से गणना के समान परिणाम मिलता है:

```
  chisq.test(observed,  correct=FALSE)#  agrees  with  our  own  calculation

          Pearson's Chi-squared test


  data:  observed

  X-squared = 44.0181, df = 1, p-value = 3.253e-11
```

R में डिफ़ॉल्ट सही है = Yates निरंतरता सुधार का उपयोग करके $\chi 2$ परीक्षण के लिए सही है (जैसा कि ऊपर prop.test() में है)। यह सुधार 95% विश्वास अंतराल को थोड़ा व्यापक बनाता है।

```
  chisq.test(observed)


          Pearson's Chi-squared test with Yates' continuity correction

  data:  observed

  X-squared = 42.2639, df = 1, p-value = 7.975e-11
```

फ़ंक्शन prop.test() इस उदाहरण में 2 x 2 आकस्मिक तालिका के लिए chsiq.test() के बराबर है:

```
  > prop.test(germinated,sown, correct = FALSE)

          2-sample test for equality of proportions without continuitycorrection

  data:  germinated out of sown

  X-squared = 44.0181, df = 1, p-value = 3.253e-11
  alternative hypothesis: two.sided

  95 percent confidence interval:
```

```
 0.3254180 0.5591974

sample   estimates:
   prop 1    prop 2

0.7500000 0.3076923


> prop.test(germinated,sown)

        2-sample test for equality of proportions with continuity
correction


data:  germinated out of sown

X-squared = 42.2639, df = 1, p-value = 7.975e-11
alternative hypothesis: two.sided

95 percent confidence interval:

 0.3165148 0.5681005

sample   estimates:
   prop 1    prop 2

0.7500000 0.3076923
```

छोटी आवृत्तियों (एक या अधिक आवृत्तियों <5) के साथ, परीक्षण आँकड़ा अच्छी तरह से $\chi 2$ वितरित नहीं है और फिशर के सटीक परीक्षण की सिफारिश की जाती है। आइए एक आवृत्ति के साथ एक उदाहरण देखें <5:

```
> sown<-c(12,13)
> germinated <-c(9,4)
> notgerminated <- sown-germinated
> observed <- matrix(c(notgerminated, germinated), ncol=2,
dimnames=list(c("species A","species B"),c("not germinated","germinated")));
observed
           not germinated germinated

species A              3          9

species B              9          4


> chisq.test(observed)

        Pearson's Chi-squared test with Yates' continuity correction

data:  observed

X-squared = 3.2793, df = 1, p-value = 0.07016


> fisher.test(observed)
```

```
        Fisher's Exact Test for Count Data

data:  observed

p-value = 0.04718

alternative hypothesis: true odds ratio is not equal to 1

95 percent confidence interval:

 0.01746573 1.11027182

sample estimates:

odds ratio
 0.1617985
```

3.4. आउटलुक: रैखिक मॉडल: (Outlook: linear models)

अधिक जटिल डेटा का विश्लेषण करने के लिए इस अध्याय में वर्णित ऐसे सरल परीक्षणों को मॉडलों द्वारा प्रतिस्थापित किया जाना है। हम एक बहुत ही सरल सामान्य रैखिक मॉडल पर एक नज़र डालेंगे जिसका उपयोग फ़र्नेस डेटा का विश्लेषण करने के लिए किया जा सकता है। दो-नमूना टी परीक्षण करने के बजाय, हम lm() फ़ंक्शन का उपयोग करके चयापचय दर की व्याख्या करने के लिए एक रैखिक मॉडल फिट कर सकते हैं:

```
> furness$Male <- as.numeric(furness$SEX == "Male")
> model <- lm(METRATE ~ Male, data = furness)
> summary(model)

Call:

lm(formula = METRATE ~ Male, data = furness)



Residuals:

     Min       1Q   Median       3Q      Max

-1037.97  -510.43   -59.37   524.33  1386.22


Coefficients:

            Estimate  Std.  Error  t  value  Pr(>|t|)
(Intercept)   1285.5       300.1   4.283  0.00106 **

Male           278.3       397.0   0.701  0.49676

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 735.2 on 12 degrees of freedom
Multiple R-squared: 0.03932,    Adjusted R-squared: -0.04073
```

2. F-statistic: 0.4912 on 1 and 12 DF, p-value: 0.4968

ध्यान दें कि हमने पुरुष फुलमार (1 = पुरुष, 0 = महिला) के लिए एक संकेतक चर बनाया है। यह मॉडल दो मापदंडों का अनुमान लगाता है। इंटरसेप्ट मादा फुलमार के लिए अनुमानित चयापचय दर है, अनुमान है कि नर नर और मादा की दर के बीच का अंतर है। फिटेड मॉडल है: 1285.5 278.3 * पुरुष = 1563.8। नर फुलमार के लिए अनुमान 278.3 जोड़ा जाता है (पुरुष = 1 से गुणा किया जाता है), मादा फुलमार के लिए इसे नहीं जोड़ा जाता है (क्योंकि नर = 0)। हमें t.test() का उपयोग करके सीधे दो चयापचय दर प्राप्त हुई:

```
t.test(METRATE ~ SEX, data = furness)


        Welch Two Sample t-test

data:  METRATE by SEX

t = -0.7732, df = 10.468, p-value = 0.4565

alternative hypothesis: true difference in means is not equal to 0

95 percent confidence interval:

 -1075.3208   518.8042

sample estimates:

mean in group Female   mean in group Male
            1285.517   1563.775
```

3.5. साहित्य: (Literature)

डालगार्ड, पी. (2008)। R न्यूयॉर्क, स्प्रिंगर के साथ परिचयात्मक सांख्यिकी।

क्विन, जी.पी. और केओफ़, एम.जे. (2002) जीवविज्ञानी के लिए प्रायोगिक डिजाइन और डेटा विश्लेषण। कैम्ब्रिज यूनिवर्सिटी प्रेस। ऑल्टमैन डी. जी

समाश्रयण विश्लेषण

आर के पॉल

भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012

ranjit.paul@icar.gov.in

1. परिचय

समाश्रयण विश्लेषण एक सांख्यिकीय विधि जिसमे दो एवं दो से अधिक मात्रात्मक चर के बीच संबंध से एक चर का मान दूसरों चरों के आपसी संबंध से ज्ञात करते हैं । इस विधि का समान्यतः उपयोग व्यपार में, सामाजिक एवं व्यवहार विज्ञान में, जैविक विज्ञान जैसे की कृषि और मत्स्य अनुसंधान इत्यादि में करतें हैं ।

दो चरों के बीच एक कार्यात्मक संबंध को एक गणितीय सूत्र द्वारा व्यक्त करतें हैं । अगर X स्वतंत्र चर एवं Y निर्भर चर है, तो एक कार्यात्मक संबंध Y = f(X) द्वारा व्यक्त करतें हैं

X के एक विशेष मान पर, फलन f, Y के अनुरूप मान को इंगित करता है ।Xऔर Y के बीच का संबंध संबंधों की प्रकृति पर निर्भर करता ह, समाश्रयण मॉडल को दो व्यापक श्रेणियां में वर्गीकृत किया जा सकता है, रेखीय समाश्रयण मॉडल एवं अरेखीय समाश्रयण मॉडल ।

2. रेखीय समाश्रयण मॉडल

हम एक बुनियादी रेखीय मॉडल पर विचार करें जिसमे केवल एक भविष्यवक्ता (predictor) चर है एवं समाश्रयण फलन रैखिक है। एक से अधिक भविष्यवक्ता चरों के साथ मॉडल निम्नानुसार दिखाया जा सकता है,

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad (1)$$

i^{th} परीक्षण में यहाँ Y_i प्रतिक्रिया चर, β_0 एवं β_1मानकें हैं, X_i, i^{th} परीक्षण में एक भविष्यवक्ता चर के मूल्य है । ε_i एक यादृच्छिक त्रुटि है जिसका माध्य शून्य एवं विचरण σ^2है ε_i एवं ε_jअसहसंबद्ध है इसीलिए सहप्रसरण 0 है ।

2.1. समाश्रयण पैरामीटर का अर्थ

समाश्रयण मॉडल (1) में β_0 और β_1मापदंडों को समाश्रयण गुणांक कहा जाता है जहाँ, β_1 समाश्रयण लाइन की ढलान है। यह X में प्रति यूनिट की बढ़ोतरी से Y के संभावना वितरण के माध्य में परिवर्तन को दर्शाता है। पैरामीटर β_0समाश्रयण लाइन Y में अन्तररोधक है ।

2.2. **Least Squares** विधि

β_0 और β_1समाश्रयण मापदंडों की सही अनुमान लगाने के लिए हम Least Squares विधि का प्रयोग करते हैं । Least Squares विधि में n squared deviationsका योग का उपयोग होता हैं । इस मापदण्ड को Qसे चिह्नित करते हैं रू

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 + \beta_1 X_i)^2 \qquad (2)$$

Least Squares विधि के अनुसार b_0 और b_1, क्रमशःβ_0 और β_1का अनुमानितमान है जोकि दिए गए अवलोकनों पर मापदण्ड Qकाकम से कम मान है ।

विश्लेषणात्मक दृष्टिकोण का प्रयोग से, हम समाश्रयण मॉडल (1) में b_0 और b_1का मान जो कि किसी विशेष नमूना आँकड़ें के सेट पर मापदण्ड Qका कम से कम मान से ज्ञात करते हैं तथा निम्नलिखित समीकरण द्वारा दिखाते हैं रू

$$\sum_{i=1}^{n} Y_i = \mathrm{nb}_0 + b_1 \sum_{i=1}^{n} X_i$$

$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2.$

इन दोनों समीकरणों को सामान्य समीकरण कहा जाता है और b_0और b_1का मान हल किया जा सकता है रू$b_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

$b_0 = \frac{1}{n}(\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i) = \overline{Y} - b\overline{\ _1X},$

जहाँ$\overline{X}$और$\overline{Y}$ए X_iऔर Y_iअवलोकनों का क्रमशः माध्य हैं।

2.3. युक्त/सज्जित समाश्रयण लाइन के गुण

एक बार मापदंडों का अनुमान प्राप्त हो तो , सज्जित लाइन होगाए

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (3)$$

$e_i = Y_i - \hat{Y}_i$जहाँ, e_iएक i^{th} रेसिडुअल/त्रुटिहै।

Least Squares विधि द्वारा अनुमानित समाश्रयण लाइन (**3**) को सज्जित करने में निम्नलिखित गुण मिलता है,

1. त्रुटियों का योग शून्य होता है,$\sum_{i=1}^{n} e_i = 0.$
2. वर्गीकृत त्रुटियों का योग,$\sum_{i=1}^{n} e_i^2$न्यूनतम होता है।
3. अवलोकित मान Y_iका योग फिट मूल्यों$\hat{Y}_i$का योग के बराबर होती है, $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$।
4. भारित त्रुटि का योग शून्य होता है, जहाँ X_i, i^{th} परीक्षण में भविष्यवक्ता चर के स्तर के आधार पर भारित मान है एवं$\hat{Y}_i$, i^{th} परीक्षण में प्रतिक्रिया चर के स्तर के आधार पर भारित मान है: $\sum_{i=1}^{n} X_i e_i = 0$और$\sum_{i=1}^{n} \hat{Y}_i e_i = 0$।

5. समाश्रयण लाइन हमेशा बिन्दु$(\overline{X}, \overline{Y})$ से गुजरता है।

2.4. पद त्रुटि के विचरण σ^2 का आकलन

Y की संभावना वितरण की परिवर्तनशीलता का एक संकेत प्राप्त करने के लिए समाश्रयण मॉडल (**1**) में त्रुटि पद के विचरण σ^2 का अनुमान लगाया जाने की जरूरतहोती है । साथ मेंए समाश्रयण फलन से संबंधित विभिन्न प्रकार का अनुमान एवं *Y* की भविष्यवाणी के अनुमान के लिएσ^2का आकलनकीआवश्यकता होती है ।जिसेए SSE = $\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} e_i^2$, द्वारा निरूपित करते हैंए जहाँ SSE त्रुटि के वर्गों का योगहै।तब त्रुटि के विचरण σ^2का आकलन निम्नलिखित के द्वारा दिया जाता हैए

$$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p},$$

जहाँ p मॉडल में शामिल मापदंडों की कुल संख्या है। हम इसे MSE के द्वारा निरूपित करते हैं ।

2.5. R^2 के द्वारा फिटिंग के माप निकलना

कई बार रेखीय एसोसिएशन की डिग्री को पता करने की जरूरत होती है । यहाँ हम एक वर्णनात्मक माप की प्रयोग करते है जो की कि मुख्यतः *Y* और *X* के बीच रैखिक एसोसिएशन की डिग्री का वर्णन करने के लिए इस्तेमाल किया जाता है । जिसे SSTO = $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$,के द्वारा इंगित करते हैं। जहाँ SSTO (total sum of squares)है । जो अवलोकनY_iमें भिन्नता की माप है । इस प्रकार SSTO यहाँ *Y* में अनिश्चितता के अनुमान की माप है जब *X* को सम्मलित नही करते हैं ।इसी तरह SSE *Y* में भिन्नता की माप है जब एक समाश्रयण मॉडल भविष्यवक्ता चर *X* के कार्यरत है ।

$$R^2 = \frac{\text{SSR}}{\text{SSTO}} = 1 - \frac{\text{SSE}}{\text{SSTO}}$$

जहाँR^2संकल्प के गुणांक कहा जाता है,$0 \leq R^2 \leq 1$. इसका मान 1 के करीब हो तो यह अधिक से अधिक Y और X के बीच रैखिक एसोसिएशन की डिग्री देता है ।

2.6. निदानिकी एवं उपचारी उपाय

जब हम एक समाश्रयण मॉडल लेते हएैं तो हम आम तौर पर यह नहीं होताहैकि अग्रिम में निश्चित हो कि लिया गया मॉडल किसी विशेष अनुप्रयोग के लिए उपयुक्त है, किसी भी मॉडल कीएक या कई गुण जैसे की समाश्रयण फलन में रैखिकता हो सकता है या त्रुटि के संदर्भ में प्रसामान्यथता ;normalityद्ध होना, विशेष रूप के डेटा के लिए उपयुक्त नहीं हो सकता है । इस भाग में हम एक मॉडल के औचित्य के अध्ययन के लिए कुछ सरल ग्राफिक तरीकों, साथ ही कुछ सुधारात्मक उपाय जब डेटा समाश्रयण मॉडल की शर्तों के अनुसार सहायक नहीं हो पर चर्चा करेंगे ।

2.7. मॉडल से प्रस्थापन का अध्ययन

हम सामान्य त्रुटियों के साथ रेखीय समाश्रयण मॉडल से प्रस्थापन के छह महत्वपूर्ण प्रकार निम्नलिखित पर विचार करेंगेरू (i) समाश्रयण फलन के रैखिकता ।

(ii) त्रुटि विचरण की स्थिरता ।

(iii) त्रुटि पदों की स्वतंत्रता।

(iv) एक या कुछ गैर अवलोकन की उपस्थिति ।

(v) त्रुटि पदों की सामान्य वितरण (normal distribution)।

(vi) एक या कई महत्वपूर्ण भविष्यवक्ता चर मॉडल से मिटाया गया हो ।

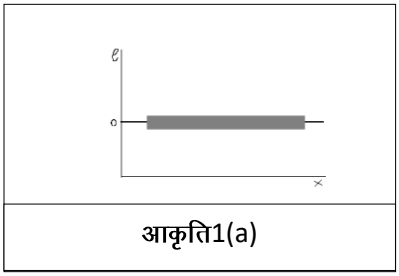
(vii) multicollinearity की उपस्थिति।

2.8. मॉडल से प्रस्थापन के लिए ग्राफिकल टेस्ट

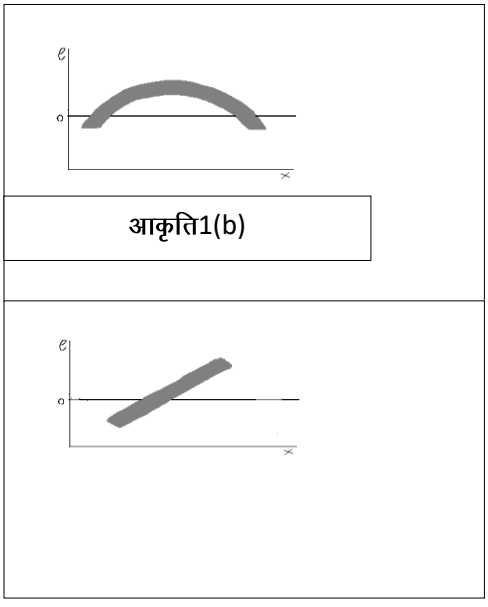
समाश्रयण मॉडल की गैर रैखिकता

आकृति 1(a) एक रेखीय समाश्रयण मॉडल का उचित प्रोटोटाइप स्थिति को दिखता है । यहाँ residualएक क्षैतिज बैंडमें है जो 0 के आसपास केंद्रित है और कोई व्यवस्थित प्रवृत्तियों के लिए सकारात्मक और नकारात्मक को प्रदर्शित नहीं करता है ।

आकृति 1(b) रेखीय समाश्रयण मॉडल से प्रस्थापन का एक प्रोटोटाइप स्थिति दिखता है जो कि एक वक्रीय समाश्रयण फलन के लिए की आवश्यकता को इंगित करता है । यहाँ residual सकारात्मक और नकारात्मक के बीच एक व्यवस्थित तरीके में बदलते हैं ।



आकृति1(a)



आकृति1(b)

आकृति1(c)	आकृति1(d)

त्रुटि विचरण की गैर स्थिरता

आकृति 1(a) में प्रोटोटाइप प्लाट दिखता है की त्रुटि पद विचरण स्थिर है, एवंआकृति1(c) में प्रोटोटाइप प्लाट दिखता है की त्रुटि पद विचरण *X* के साथ बढ़ती जाती है जो की "Megaphone" प्रकार दीखता है ।

Outliers की उपस्थिति

Outliers चरम अवलोकन है । अवशिष्ट (**residual**) outliers**, residual** प्लाट *X* या$\hat{Y}$के खिलाफ से पहचाना जा सकता है।

त्रुटिपदों के गैर स्वतंत्रता आकृति 1(d) में एक प्रोटोटाइप अवशिष्ट प्लाट समय संबंधित प्रवृत्ति असर दिखाता है, जो एक रेखीय समय संबंधित प्रवृत्ति प्रभाव का चित्रण है।

त्रुटि पदों के गैर **Normality**

त्रुटि पदों के Normality का अध्ययन विभिन्न ग्राफिक तरीकों से residuals परीक्षण से किया जा सकता है । जैसे की आवृत्तियों की तुलना एवं सामान्य संभावना प्लाट।

महत्वपूर्ण कारक चर की अनुपस्थिति

अतिरिक्त भविष्यवक्ता चर के खिलाफ residual plot से यह पता करते है की अतिरिक्त भविष्यवक्ता चर के विभिन्न स्तर के साथ व्यवस्थित ढंग से भिन्न residualtrend है या नहीं ।

2.9. मॉडल से प्रस्थापन के लिए सांख्यिकीय परीक्षण

Randomness के लिए टेस्ट

Durbin-Watson test: अगरρ,autocorrelation गुणांक है और $e_t = Y_t - \hat{Y}_t$तो

$H_0: \rho = 0$

$H_1: \rho > 0$

$$D = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{n\sum_{t=1} e_t^2}$$

निष्कषः *D* के छोटे मान से यह पता चलता है कि ρ>0।

Normality के लिए परीक्षण

Normality के लिए **Correlation** परीक्षणः Kolmogorov-Smirnov testऔरAnderson-Darling Test.

Tests for Constancy of Error Variance: Modified Levene Test और White Test

Outlying अवलोकनों के लिए परीक्षणः

(i)**Elements of Hat Matrix:** $H = X(X'X)^{-1}X'$, जहाँ **X** व्याख्यात्मक चर के लिए मैट्रिक्स है, जो बड़े मानों के आँकड़े बिंदुए outliers को दर्शाते हैं ।

(ii)***WSSD$_i$***: x-space में किसी दूरदराज के बिंदु का पता लगाने के लिए *WSSD$_i$* का उपयोग करते हैं ।

(iii) ***Cook's D$_i$:*** Cook's D$_i$$\hat{y}$में परिवर्तन को मापता है जब ithअवलोकन मापदंडों के आकलन में इस्तेमाल नहीं किया गया हो ।

(iv) ***DFFITS$_i$:*** *DFFIT* का उपयोग $\left(\hat{y} - \hat{y}_{(i)}\right)$के ithघटक में अंतर मापने में होता है ।

(v)$\text{DFBETAS}_{j(i)}$: व्यक्तिगत समाश्रयण गुणांक के लिए प्रभावशाली अवलोकनों $\text{DFBETAS}_{j(i)}, j = 1,2,\ldots, p+1$, द्वारा पहचाने जाते हैं ।

(vi) COVRATIO_i***:*** अनुमानित समाश्रयण गुणांक के विचरण–सहप्रसरण मैट्रिक्स पर *i*thअवलोकन के प्रभाव को दो विचरण–सहप्रसरण matrices के निर्धारकों के अनुपात से मापा जाता है। इस प्रकार, COVRATIO समाश्रयण गुणांक के अनुमानों की शुद्धता पर*i*thअवलोकन के प्रभाव को दर्शाता है।

(vii) FVARATIO_i***:*** जब एक अवलोकन हटाया जाता है तो$\hat{y}_i$के विचरण में परिवर्तन का पता FVARATIO_i से लगाते है ।

2.10. Multicollinearity

Multicollinearityकी समस्याओं को निम्नलिखित तरीकों से दूर कर सकतें हैं -

अतिरिक्त डेटा का संग्रह, Model respecification और Ridge Regression.

उदाहरण:

तालिका 1

Case	X_{11}	X_{21}	X_{31}	Y_i
1	12.980	0.317	9.998	57.702
2	14.295	2.028	6.776	59.296
3	15.531	5.305	2.947	56.166
4	15.133	4.738	4.201	55.767
5	15.342	7.038	2.053	51.722
6	17.149	5.982	-0.055	60.446
7	15.462	2.737	4.657	60.715
8	12.801	10.663	3.048	37.447
9	17.039	5.132	0.257	60.974
10	13.172	2.039	8.738	55.270
11	16.125	2.271	2.101	59.289

12	14.340	4.077	5.545	54.027
13	12.923	2.643	9.331	53.199
14	14.231	10.401	1.041	41.896
15	15.222	1.220	6.149	63.264
16	15.740	10.612	-1.691	45.798
17	14.958	4.815	4.111	58.699
18	14.125	3.153	8.453	50.086
19	16.391	9.698	-1.714	48.890
20	16.452	3.912	2.145	62.213
21	13.535	7.625	3.851	45.625
22	14.199	4.474	5.112	53.923
23	15.837	5.753	2.087	55.799
24	16.565	8.546	8.974	56.741
25	13.322	8.589	4.011	43.145
26	15.949	8.290	-0.248	50.706

तलिका 2:प्रभावशाली प्रेक्षणों के संकेतक

Case	r_i	t_i	t_I*=s.t/s_i	h_{ii}	D_i	$WSSD_i$
1	0.460	0.289	0.281	0.215	0.005	39*
2	1.253	0.732	0.724	0.093	0.013	12
3	0.377	0.215	0.210	0.048	0.001	1
4	0.044	0.025	0.026	0.042	0.000	1
5	-0.256	-0.146	-0.141	0.053	0.000	3
6	1.010	0.611	0.602	0.155	0.017	20
7	0.389	0.226	0.221	0.081	0.001	7
8	0.132	0.088	0.086	0.301	0.001	41

9	0.432	0.262	0.256	0.155	0.003	18
10	0.589	0.355	0.347	0.147	0.005	23
11	-3.302	-2.021	-2.193	0.173	0.214	14
12	-0.406	-0.232	-0.226	0.053	0.001	3
13	0.194	0.118	0.117	0.163	0.001	24
14	-0.268	-0.164	-0.161	0.175	0.001	23
15	0.802	0.476	0.469	0.122	0.007	15
16	-0.482	-0.295	-0.289	0.177	0.005	26
17	3.756	2.134	2.343	0.041	0.048	0
18	-6.072	-3.589	-5.436	0.114	0.412	8
19	-1.198	-0.727	-0.719	0.160	0.025	24
20	1.126	0.666	0.658	0.114	0.014	11
21	0.449	0.266	0.259	0.119	0.003	12
22	0.791	0.453	0.444	0.055	0.003	3
23	-0.060	-0.035	-0.032	0.059	0.000	3
24	0.574	1.181	1.188	0.927	4.409	19
25	0.268	0.163	0.158	0.159	0.001	19
26	-0.606	-0.356	-0.350	0.101	0.004	11

तलिका 3:प्रभावशाली प्रेक्षणों के संकेतक

Case	Cov Ratio	Dffits	Intercept	X1	X2	X3
				DFBETAS		
1	1.512	0.148	0.056	-0.053	-0.006	0.006
2	1.203	0.232	0.062	-0.042	-0.042	-0.050
3	1.254	0.047	-0.005	0.010	-0.008	-0.007

4	1.257	0.005	0.000	0.000	-0.001	0.000
5	1.267	-0.033	-0.001	-0.001	-0.006	0.006
6	1.331	0.258	-0.095	0.132	-0.042	-0.050
7	1.299	0.068	-0.005	0.015	-0.036	-0.005
8	1.721	0.057	0.027	-0.034	0.026	-0.006
9	1.408	0.109	-0.030	0.048	-0.035	-0.031
10	1.380	0.144	0.058	-0.058	-0.041	0.016
11	0.639	-1.004	-0.154	-0.045	0.776	0.525
12	1.260	-0.054	-0.017	0.014	0.014	0.000
13	1.435	0.051	0.017	-0.19	-0.004	0.013
14	1.452	-0.074	-0.026	0.031	-0.35	0.015
15	1.315	0.175	-0.008	0.033	-0.105	0.002
16	1.441	-0.134	-0.014	0.014	-0.044	0.047
17	0.496	0.482	0.061	-0.17	-0.107	-0.046
18	0.410	-1.945	0.362	-0.308	-0.220	-1.177
19	1.301	-0.341	0.031	-0.045	-0.080	0.094
20	1.252	0.236	-0.055	0.097	-0.105	-0.051
21	1.350	0.095	0.054	-0.061	0.024	-0.018
22	1.228	0.108	0.052	-0.048	-0.028	-0.020
23	1.279	-0.008	0.001	-0.002	0.001	0.002
24	12.715	4.230	-3.642	3.276	3.180	3.934
25	1.426	0.069	0.031	-0.039	0.029	-0.003
26	1.309	-0.117	0.000	-0.007	-0.016	0.043

तलिका 4:समाश्रयण गुणांक और सारांश आँकड़े

Description	b_0	b_1	b_2	b_3	s	R^2	Max VIF	Min e.v.	Max R_i^2
All Data (n=26)	8.11	3.56	-1.63	0.34	1.80	0.94	2.82	0.210	0.65
Delete (11, 17, 18)	7.17	3.66	-1.79	0.40	0.51	0.99	2.85	0.210	0.65
Delete (24)	30.91	2.39	-2.14	-0.36	1.78	0.94	30.64	0.017	0.97
Delete (11, 17, 18, 24)	24.27	2.79	-2.11	-0.16	0.50	0.99	171.90	0.003	0.99
Ridge k=0.05 (n=22)	14.28	3.22	-1.73	0.25	0.66	0.99	10.20	0.053	0.90
Delete X3 (n=22)	19.50	3.03	-2.00		0.49	0.99	1.02	0.863	0.02

कुछ चयनित संदर्भ

वेल्स्ली, डी.ए., कुह, ई एवं वेल्श, आर.ई. (2004): "समाश्रयण निदान – प्रभावशाली डेटा एवं collinearity, के स्रोतों की पहचान", न्यू यॉर्क.: विले लिमिटेड ।

बार्नेट, वी और लुईस, टी (1984): "सांख्यिकीय डेटा में outliers", न्यू यॉर्क: विले लिमिटेड ।

चटर्जी, एस और मूल्य, बी (1977): "उदाहरण के द्वारा समाश्रयण विश्लेषण", न्यू यॉर्क: जॉन विले एंड संस ।

ड्रेपर, एन.आर. एवं स्मिथ, एच (1998): "एप्लाइड समाश्रयण विश्लेषण", न्यू यॉर्क: विले पूर्वी लिमिटेड ।

क्लेंबौम, डी.जी. ,वं कुप्पेर, एल.एल. (1978): "एप्लाइड समाश्रयण विश्लेषण और अन्य बहुभिन्नरूपी चर तकनीक", मैसाचुसेट्सः देक्स्वरी प्रेस ।

मांटगोमेरी, डी.सी., पेक, एवं भिनिंग, जी (2003): "रेखीय समाश्रयण विश्लेषण का परिचय ", तीसरा संस्करण, न्यू यॉर्क: जॉन विले एंड संस ।

सांख्यिकीय आनुवंशिकी में गैर पैरामीट्रिक तरीके

हिमाद्री शेखर रॉय
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
himadri.roy@icar.gov.in

परीक्षण (आमतौर पर 'परिकल्पना परीक्षण' कहा जाता है) सांख्यिकीय जांच में एक प्रमुख भूमिका निभाते हैं। सांख्यिकीय परीक्षण सत्य या अन्यथा, परिकल्पना की जांच करने से संबंधित हैं। (धारणाएं, दावे, अनुमान, आदि) एक या अधिक पापूलेशन की कुछ विशेषताओं के बारे में लगभग सभी बड़े और छोटे नमूना परीक्षण जैसे t, F और χ^2 इस धारणा पर आधारित हैं कि मूल जनसंख्या (जिससे नमूना लिया गया है) का एक विशिष्ट वितरण होता है, जैसे सामान्य वितरण। वितरण को आमतौर पर कुछ मापदंडों के माध्यम से परिभाषित किया जाता है। परीक्षणों को ऐसी धारणा की आवश्यकता नहीं होती है। इसलिए गैर-पैरामीट्रिक परीक्षणों को . के रूप में भी जाना जाता है वितरण मुक्त परीक्षण। गैर-पैरामीट्रिक शब्द इस तथ्य को संदर्भित करता है कि कोई पैरामीटर नहीं हैं आम तौर पर इस्तेमाल किए जाने वाले पैरामीटर शब्द के पारंपरिक अर्थ में शामिल है। आँकड़े नमूना डेटा के कुछ सरल पहलुओं का उपयोग करते हैं जैसे माप के संकेत, क्रमसंबंध या श्रेणी आवृत्तियों। इसलिए, स्केल को खींचने या संपीड़ित करने से नहीं होता है उन्हें बदलो। परिणामस्वरूप, गैर-पैरामीट्रिक परीक्षण आँकड़ों का शून्य वितरण हो सकता है मूल जनसंख्या वितरण के आकार की परवाह किए बिना निर्धारित किया जाता है।

पैरामीट्रिक परीक्षणों जैसे t, F और $\chi 2$ के आधार पर परीक्षणों से निकाले गए निष्कर्ष शायद माता-पिता की जनसंख्या का वितरण सामान्य नहीं होने पर गंभीर रूप से प्रभावित होता है। ये प्रभाव हो सकते हैं अधिक हो जब नमूना आकार छोटा हो। इस प्रकार जब वितरण के बारे में संदेह होता है मूल पापूलेशन, एक गैर-पैरामीट्रिक पद्धति का उपयोग किया जाना चाहिए। कई स्थितियों में विशेष रूप से सामाजिक और व्यवहार विज्ञान के अवलोकन संख्यात्मक पर लेना मुश्किल या असंभव है संख्यात्मक पैमाने पर गैर-पैरामीट्रिक परीक्षण ऐसी स्थितियों के लिए उपयुक्त हैं।

सांख्यिकीय परीक्षण में पहला कदम एक परिकल्पना तैयार करना है। एक परिकल्पना के बारे में एक कथन पापूलेशन है। पापूलेशन से नमूने द्वारा प्राप्त जानकारी के आधार पर इसकी संभाव्यता का मूल्यांकन किया जाता है। एक परीक्षण में आम तौर पर दो परिकल्पनाएं शामिल होती हैं। एक

के बारे में एक दावा 'मौजूदा' स्थितियों के पक्ष में जनसंख्या को शून्य परिकल्पना के रूप में लिया जाता है और इसे भ्0 के रूप में दर्शाया जाता है। शून्य परिकल्पना के निषेध को वैकल्पिक परिकल्पना के रूप में जाना जाता है और इसे भ्1 के रूप में दर्शाया जाता है। एच 1 किसी परीक्षण को एक तरफा या दो तरफा के रूप में वर्गीकृत करने में निर्णायक भूमिका निभाता है। हम पहले एक आँकड़ा विकसित करते हैं टी (कहते हैं) नमूना टिप्पणियों के आधार पर। आँकड़ा ज् तय करता है कि अस्वीकार करना है या शून्य परिकल्पना को स्वीकार करें। आमतौर पर ज् कुछ वितरण का अनुसरण करता है। इस वितरण के आधार पर टी की सीमा दो समूहों में विभाजित है; क्रान्तिक क्षेत्र और स्वीकृति का क्षेत्र। अगर नमूना बिंदु क्रान्तिक क्षेत्र में आता है, हम शून्य परिकल्पना को अस्वीकार करते क्रान्तिक क्षेत्र का आकार उस जोखिम पर निर्भर करता है जिसे हम स्वीकार करना चाहते हैं जो अंततः परीक्षण का महत्व स्तर देता है। यह α द्वारा निरूपित है । यह शून्य परिकल्पना को अस्वीकृत करने की प्रायिकता को भी दर्शाता है जब यह त्रिशंकु होती है टाइप प् त्रुटि के रूप में जाना जाता है। टाइप प्प् त्रुटि भ्0 को अस्वीकार करने की संभावना है जब भ्1 सत्य है और β द्वारा निरूपित। आमतौर पर इस्तेमाल किए जाने वाले महत्व के स्तर 5þ और 1þ (α = .05 और .01) हैं। अंत में, ए निष्कर्ष ज् के महत्वपूर्ण क्षेत्र में गिरने या न गिरने के मान के आधार पर निकाला जाता है। कुछ सामान्य रूप से उपयोग किए जाने वाले गैर-पैरामीट्रिक परीक्षणों की अगली कड़ी में चर्चा की गई है।

1. यादृच्छिकता के लिए रन परीक्षण

रन टेस्ट का उपयोग यह जांचने के लिए किया जाता है कि अवलोकनों का एक सेट यादृच्छिक है या नहीं एक अनंत पापूलेशन से नमूना। यादृच्छिकता के लिए परीक्षण का बहुत महत्व है क्योंकि यादृच्छिकता की धारणा सांख्यिकीय अनुमान का आधार है। इसके अलावा, यादृच्छिकता के लिए परीक्षण हैं समय श्रृंखला विश्लेषण के लिए महत्वपूर्ण। यादृच्छिकता से प्रस्थान कई रूप ले सकता है।

H_0: नमूना मान एक यादृच्छिक अनुक्रम से आते हैं

H_1: नमूना मान एक गैर-यादृच्छिक अनुक्रम से आते हैं

टेस्ट आँकड़ाः मान लें कि r रनों की संख्या है (एक रन एक ही तरह के बंधे हुए चिन्ह का एक क्रम है अन्य प्रकार के संकेतों द्वारा)। रनों की संख्या ज्ञात करने के लिए, प्रेक्षणों को उनके घटना का क्रम को सूचीबद्ध किया गया है । प्रत्येक प्रेक्षण को '+' चिह्न द्वारा दर्शाया जाता है यदि यह पिछले प्रेक्षण से अधिक है अवलोकन और '-' चिह्न द्वारा यदि यह पिछले अवलोकन से कम है। रनों की कुल संख्या ऊपर (+s) और नीचे (-) गिना जाता है। बहुत कम रन इंगित करते हैं कि अनुक्रम यादृच्छिक नहीं है (है दृढ़ता) और बहुत अधिक रन यह भी संकेत करते हैं कि अनुक्रम यादृच्छिक नहीं है (झिगज़ैग है)।

महत्वपूर्ण मूल्यः परीक्षण के लिए महत्वपूर्ण मूल्य तालिका से n और (α) महत्व के वांछित स्तर पर दिए गए मान के लिए प्राप्त किया जाता है । माना यह मान तब r_c.

निर्णय नियमः यदि तब r_c . (निचला)) $\leq r \leq r_c$ (ऊपरी) H_0 स्वीकार करते हैं। अन्यथा H_0 को अस्वीकार करें।

बंधित मानः यदि कोई प्रेक्षण अपने पूर्ववर्ती प्रेक्षण के बराबर है तो उसे शून्य से निरूपित करें। जबकि रनों की संख्या की गणना करते हुए इसे अनदेखा करें और तदनुसार n के मान को कम करें।

बड़े नमूना आकारः जब नमूना आकार 25 से अधिक होता है तो महत्वपूर्ण मूल्य rc प्राप्त किया जा सकता है एक सामान्य वितरण सन्निकटन का उपयोग करना।

महत्व के 5% स्तर पर दो तरफा परीक्षण के लिए महत्वपूर्ण मान है

$$r_c\,(\text{lower}) = \mu - 1.96\ \sigma \ ; \ r_c\,(\text{upper}) = \mu + 1.96\ \sigma$$

ये हैं एकतरफा परीक्षणों के लिए

r_c (left tailed) = μ - 1.65 σ, if $r \leq r_c$, reject H_0

r_c (right tailed) = μ + 1.65 σ, if $r \geq r_c$, reject H_0

where $\boldsymbol{\mu = \frac{2n-1}{3}}$ and $\boldsymbol{\sigma = \sqrt{\frac{16n-29}{90}}}$

उदाहरण 1:

यूके से चयनित कृषि उत्पादन इनपुट के आयात के मूल्य पर डेटा हाल के 12 वर्षों के दौरान एक काउंटी (मिलियन डॉलर में) नीचे दिया गया हैः क्या अनुक्रम यादृच्छिक है?

5.2 5.5 3.8 2.5 8.3 2.1 1.7 10.0 10.0 6.9 7.5 10.6

H_0: अनुक्रम यादृच्छिक है।

H_1: अनुक्रम यादृच्छिक नहीं है।

5.2	5.5	3.8	2.5	8.3	2.1	1.7	10.0	10.0	6.9	7.5	10.6
	+	-	-	+	-	-	+	0	-	+	+

यहां n = 11, रन की संख्या r = 7. α = 5% (दो तरफा परीक्षण) के लिए महत्वपूर्ण n मान तालिका r_c (निचला) = 4 और r_c (ऊपरी) = 10 है। चूंकि r_c (निचला) $\leq r \leq r_c$ (ऊपरी), यानी, देखा गया है r 4 और 10 के बीच स्थित है, H0 स्वीकार किया जाता है। क्रम यादृच्छिक है।

2. वाल्ड-**वुल्फोविट्ज टू-सैंपल रन टेस्ट**

वाल्ड-वुल्फोविट्ज रन टेस्ट का उपयोग यह जांचने के लिए किया जाता है कि क्या दो यादृच्छिक नमूने समान वितरण वाली पापूलेशन आते हैं । यह परीक्षण औसत या प्रसार में अंतर का पता लगा सकता है या दो पापूलेशन के बीच कोई अन्य महत्वपूर्ण पहलू। यह परीक्षण तब प्रभावी होता है जब प्रत्येक नमूना आकार मध्यम रूप से बड़ा है (10 से अधिक या उसके बराबर)।

***H*₀**: दो नमूने समान वितरण वाली पापूलेशन से आते हैं

***H*₁**: दो नमूने अलग-अलग वितरण वाली पापूलेशन से आते हैं

टेस्ट आँकड़ाः मान लीजिए *r* रनों की संख्या को दर्शाता है। *r* प्राप्त करने के लिए से $n_1 + n_2$ प्रेक्षणों को सूचीबद्ध कीजिए परिमाण के क्रम में दो नमूने। एक नमूने के प्रेक्षणों को *x* और दूसरे *y* के द्वारा प्रेक्षणों को निरूपित करें । रनों की संख्या गिनें।

महत्वपूर्ण मूल्यः स्थान के अंतर से कुछ रन बनते हैं और प्रसार में अंतर का भी परिणाम कुछ ही रनों में होता है । नतीजतन, इस परीक्षण के लिए महत्वपूर्ण क्षेत्र हमेशा एकतरफा होता है। छै रनों की संख्या कम है या नहीं, यह तय करने के लिए महत्वपूर्ण मान तालिका से प्राप्त किया जाता है। तालिका 5% स्तर पर n1 (नमूना 1 का आकार) और n2 (नमूना 2 का आकार) के लिए महत्वपूर्ण मान r_c देती है महत्व का।

निर्णय नियमः यदि r $r \leq r_c$ H_0. को अस्वीकार करता है।

टाईः यदि *x* और y प्रेक्षणों का मान समान है तो प्रेक्षण *x*(*y*)) को पहले रखें यदि *x*(*y*)) का रन हो अवलोकन जारी है।

बड़े नमूना आकारः 20 महत्वपूर्ण मूल्य से बड़े नमूने के आकार के लिए r_c नीचे दिया गया है।

$$r_c = \mu - 1.96\,\sigma \text{ at 5\% level of significance}$$

$$\text{where } \mu = 1 + \frac{2n_1n_2}{n_1+n_2} \text{ and } \sigma = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}$$

उदाहरण 2: यह निर्धारित करने के लिए कि क्या एक नया संकर बीज बोने से झाड़ीदार फूल वाला पौधा पैदा होता है,

निम्नलिखित डेटा एकत्र किया गया था। जांच परीक्षण करें कि क्या डेटा इंगित करता है कि नया संकर बड़ा उत्पादन करता है
वर्तमान किस्म की तुलना में झाड़ियाँ?
झाड़ियों का घेरा (इंच में)

हाइब्रिड	***x***	31.8	32.8	39.2	36.0	30.0	34.5	37.4
वर्तमान	***y***	35.5	27.6	21.3	24.8	36.7	30.0	
विविधता								

Ho : एक्स और वाई आबादी समान हैं
H_1 : और ल झाड़ियों की परिधि में कुछ अंतर है।

संयुक्त आदेशित डेटा पर विचार करें।

21.3	24.8	27.6	30.0	30.0	31.8	32.8	34.5	35.5	36.0	36.7	37.4	39.2
y	***y***	***y***	***y***	***x***	***x***	***x***	***x***	***y***	***x***	***y***	***x***	***x***

टेस्ट आँकड़ा r = 6 (रनों की कुल संख्या)। n_1 = 7 और n_2 = 6 के लिए, r_c 5 % स्तर पर क्रांतिक मान आरसी महत्व 3 है। चूंकि तझ तब, हम Ho को स्वीकार करते हैं कि **x** और **y** का वितरण समान है।

3. दो नमूनों के लिए माध्यिका परीक्षण:

यह टेस्ट करने के लिए कि एक ही पापूलेशन से दो नमूने आते हैं या नहीं, माध्यिका परीक्षण का उपयोग किया जाता है। यह अधिक है रन टेस्ट की तुलना में कुशल लेकिन प्रत्येक नमूना कम से कम 10 आकार का होना चाहिए। इस मामले में,

परीक्षण की जाने वाली परिकल्पना है

H0: दो नमूने समान वितरण वाली पापूलेशन से आते हैं।

H1 : दो नमूने अलग-अलग वितरण वाली पापूलेशन से आते हैं।

टेस्ट आँकड़ाः χ^2 (काई-स्क्वायर)

परीक्षण आँकड़ों के मूल्य का परीक्षण करने के लिए n1 और n2 आकार के दो नमूने संयुक्त हैं। आकार n = n1+n2 के संयुक्त नमूने का माध्यिका M प्राप्त होता है। की संख्या प्रत्येक नमूने के लिए माध्यिका M के नीचे और ऊपर के प्रेक्षणों का निर्धारण किया जाता है। यह तब है नीचे दिए गए तरीके से 2×2 आकस्मिक तालिका के रूप में विश्लेषण किया गया।

	अवलोकनों की संख्या		
	नमूना 1	नमूना 2	कुल
माध्यिका से ऊपर	**a**	**b**	**a+b**
माध्यिका के नीचे	**c**	**d**	**c+d**
	$a+c=n_1$	**$b+d=n_2$**	**$n=a+b+c+d$**

$$\textbf{Test Statistic: } \chi^2 = \frac{(ad-bc)^2\,(a+b+c+d)}{(a+c)(b+d)(a+b)(c+d)}$$

निर्णय नियमः यदि $\chi^2 \geq \chi^2$, H_0 को अस्वीकार करें अन्यथा इसे स्वीकार करें।

टाईः संबंधों को नजरअंदाज कर दिया जाता है और द को तदनुसार समायोजित किया जाता है।

नोटः इस परीक्षण को ा नमूनों तक बढ़ाया जा सकता है। नीचे और ऊपर प्रेक्षणों की संख्या $2 \times k$ आकस्मिक तालिका से संयुक्त माध्य M .

उदाहरण 3 : परीक्षण के लिए उदाहरण 1 की समस्या पर एक माध्यिका परीक्षण करें कि दोनों नमूने एक ही पोपुलेसन से आते हैं।

H_0 : x और y समष्टि समरूप हैं।

H_1: x और y झाड़ियों की परिधि में कुछ अंतर है।

सातवां मान 32.8 संयुक्त क्रमित अनुक्रम की माध्यिका है।

	अवलोकनों की संख्या नमूना 1	नमूना 2	कुल
	x	*y*	
Above M	4	2	6
Below M	2	4	6
	6	6	12

$$\chi^2 = \frac{12(16-4)^2}{6.6.6.6} = \frac{4}{3} = 1.33 .$$

चूंकि $\chi^2 = 1.33 < \chi_c^2 = 3.84$, Ho स्वीकार किया जाता है। यह निष्कर्ष निकाला गया है कि दो नमूने से आते हैं एक ही पापूलेशन । संकर और वर्तमान किस्म के परिधि में झाड़ी का। कोई महत्वपूर्ण अंतर नहीं है

नोटः यह उदाहरण परीक्षण प्रक्रिया को प्रदर्शित करने के लिए सरल है। वास्तविक स्थिति में द को पर होना चाहिए कम से कम 20 और प्रत्येक सेल आवृत्ति कम से कम 5.

प्प्प्-61

4. **सुमेलित जोड़ियों के लिए साइन टेस्ट :**

कई स्थितियों में, दो उपचारों के प्रभाव की तुलना रुचि का है लेकिन अवलोकन जोड़े में होता है। इस प्रकार दो नमूने वास्तव में यादृच्छिक नहीं हैं। ऐसी जोड़ी के कारण निर्भरता सामान्य दो नमूना परीक्षण उपयुक्त नहीं हैं। ऐसी स्थितियों में जब एक युग्म का सदस्य उपचार । से जुड़ा है और दूसरा उपचार ठ से, गैर-पैरामीट्रिक साइन टेस्ट की व्यापक प्रयोज्यता है। गुणात्मक डेटा होने पर भी इसे लागू किया जा सकता है उपलब्ध हैं। जैसा कि नाम से पता चलता है कि यह प्रतिक्रिया अंतर के संकेतों पर आधारित है क्प. अगर पजी प्रेक्षणों के जोड़े (x_i, y_i) । तब $Di = x_i - y_i$। परीक्षण की जाने वाली परिकल्पना है -

H_0 : उपचार ए और बी के प्रभाव में कोई अंतर नहीं है।

H_1 : A, B से बेहतर है

परीक्षण आँकड़ाः मान लीजिए **S** '-' चिह्नों की संख्या है।

क्रिटिकल वैल्यूः क्रिटिकल वैल्यू S_c, द के अनुरूप जोड़े की संख्या तालिका 3 में दी गई है। महत्व स्तर α_1 द्वारा दिया जाता है क्योंकि महत्वपूर्ण क्षेत्र एक तरफा (बाएं पूंछ वाला) है।

निर्णय नियमः **यदि** $S \leq S_c$, H_0 को अस्वीकार करता है, अन्यथा भ्0 को स्वीकार करता है।

टाईः यदि एक जोड़ी के दो मान बराबर हैं, तो उस जोड़ी को अस्वीकार कर दें और की संख्या कम करें तदनुसार अवलोकन।

नोटः यदि वैकल्पिक H_1 यह है कि और A और B के प्रभाव में कुछ अंतर है, तो S दर्शाता है

या तो नकारात्मक संकेतों की संख्या या सकारात्मक संकेतों की संख्या जो भी हो छोटा। एक महत्वपूर्ण क्षेत्र दो तरफा है और S_c को खोजने के लिए महत्वपूर्ण स्तर α_2 द्वारा दिया जाता है।

उदाहरण 4 :

एक बाजार अध्ययन में, नींबू पानी के दो ब्रांडों की तुलना की गई। 50 न्यायाधीशों में से प्रत्येक निम्नलिखित परिणामों के साथ दो नमूनों का स्वाद चखा, एक ब्रांड A और दूसरा ब्रांड B। 35 पसंदीदा ब्रांड ए, 10 पसंदीदा बी, और 5 अंतर नहीं बता सके। इस प्रकार एन = 45 और एस = 10। चूंकि $S < Sc$, हम ब्रांड A को पसंद किए जाने वाले विकल्प H_1 के पक्ष में बिना किसी अंतर के H_0 को अस्वीकार करते हैं वैकल्पिक भ्1 के पक्ष में अंतर है कि ब्रांड । को प्राथमिकता दी जाती है।

5. मिलान जोड़े के लिए विलकॉक्सन ने रैंक टेस्ट पर हस्ताक्षर किए

उन स्थितियों में जहां दो नमूनों में टिप्पणियों के बीच किसी प्रकार का युग्म होता है साधारण दो नमूना परीक्षण उपयुक्त नहीं हैं। ऐसी स्थितियों में हस्ताक्षरित रैंक परीक्षण उपयोगी होते हैं। जब अवलोकनों को डेटा मापा जाता है, तो साइन टेस्ट की तुलना में हस्ताक्षरित रैंक परीक्षण अधिक कुशल होता है क्योंकि यह देखे गए अंतरों के परिमाण को ध्यान में रखता है, यदि के बीच का अंतर दो उपचारों ए और बी की प्रतिक्रिया का परीक्षण किया जाना है परीक्षण परिकल्पना है

H_0 : उपचार ए और बी के प्रभाव में कोई अंतर नहीं है।

H_1 : उपचार A, B से बेहतर है।

टेस्ट आँकड़ाः **T** नकारात्मक संकेतों के साथ रैंकों के योग का प्रतिनिधित्व करता है। **T** की गणना के लिए, प्राप्त करें अंतर $Di = xi - yi$ लप जहां जहां x_i, उपचार A की प्रतिक्रिया है और y_i का उपचार

B । मतभेदों के निरपेक्ष मूल्यों को रैंक करें। सबसे छोटा रैंक 1 दें। संबंधों को औसत दिया जाता है रैंक। देखे गए अंतर के प्रत्येक रैंक चिह्न को असाइन करें। नकारात्मक रैंकों का योग प्राप्त करें।

महत्वपूर्ण मानः तालिका 4 में द संख्या के लिए $\mathbf{T_c}$ दिया गया है। जोड़े का। महत्व स्तर α_1 द्वारा दिया गया है महत्वपूर्ण क्षेत्र एक तरफा है।

निर्णय नियमः $T \leq T_c$ अस्वीकार H_0, अन्यथा इसे स्वीकार करें।

टाईः उस जोड़ी को छोड़ दें जिसके लिए अंतर = 0 है और तदनुसार द घटाएं। समान अंतर हैं

औसत रैंक सौंपा।

उदाहरण 5: कम करने के लिए दवा से पहले और बाद में दस रोगियों का रक्तचाप पढ़ना रक्तचाप इस प्रकार है। विकल्प के खिलाफ कोई प्रभाव नहीं की शून्य परिकल्पना का परीक्षण करें वह दवा प्रभावी है।

रोगी		1	2	3	4	5	6	7	8	9	10
पहले इलाज	*x*	86	84	78	90	92	77	89	90	90	86
बाद में इलाज	*y*	80	80	92	79	92	82	88	89	92	83
मतभेद		6	4	-14	11	0	-5	1	1	-2	3
रैंक		7	5	9	8	छोड़ें	6	1.5	1.5	3	4
साइन		+	+	-	+	छोड़ें	-	+	+	-	+

ऋणात्मक अंतरों का रैंक योग = 3+6+9 = 18. इसलिए परीक्षण आँकड़ों का मान $\mathbf{T}$ = 18. n के लिए = 9 और α_1 = 5% T_c= 8 तालिका 4 से। चूंकि $\mathbf{T>T_c}$ ज्झज्ब दवा के कोई प्रभाव नहीं होने की अशक्त परिकल्पना है स्वीकार किया।

6. कोलमोगोरोव-स्मिरनोव टेस्ट

उन स्थितियों में जहां दो नमूनों में असमान संख्या में अवलोकन हैं, कोलमोगोरोव स्मिरनोव परीक्षण उपयुक्त है। इस परीक्षण का उपयोग यह जांचने के लिए किया जाता है कि क्या कोई महत्व है दो उपचारों ए और बी के बीच अंतर (मान लीजिए)। परीक्षण परिकल्पना है

H_0: उपचार ए और बी के प्रभाव में कोई अंतर नहीं है।

H_1 : उपचार ए और बी के प्रभाव में कुछ अंतर है।

परीक्षण आँकड़ाः परीक्षण आँकड़ा है $D_{m,n}$ =sup $|F_m(x) - G_n(x)|$, F और G नमूने हैं क्रमशः दो नमूनों के नमूना अवलोकनों का प्रयोगाश्रित वितरण संबंधित नमूना आकार एम और एन। $F(x_i)$ की गणना के नमूना प्रेक्षणों की औसत संख्या के रूप में की जाती है पहला नमूना जो x_i से कम है। इसी प्रकार $G(x_i)$ की गणना की जाती है। $D_{m,n}$ एन का सबसे बड़ा मूल्य है F(x) और G(x) के बीच पूर्ण अंतर।

महत्वपूर्ण मानः $D_{m,n}$, का सारणीबद्ध मूल्य एम, एन और के लिए विभिन्न मूल्यों के लिए उपलब्ध है महत्व के विभिन्न स्तर। तालिका 4 में द संख्या के लिए दिया गया है जोड़े का। महत्व स्तर है महत्वपूर्ण क्षेत्र के रूप में α_1 द्वारा दिया गया एक तरफा है।

निर्णय नियमः यदि $D_{m,n}$ का परिकलित मूल्य $D_{m,n}$ के सारणीबद्ध मूल्य से अधिक है, तो H_0 को अस्वीकार कर दिया जाता है अन्यथा इसे स्वीकार कर लिया जाता है।

उदाहरण 6: निम्नलिखित डेटा अलग-अलग बैटरी के जीवनकाल (घंटे) का प्रतिनिधित्व करते हैं

ब्रांड A 40 30 40 45 55 30

ब्रांड B 50 50 45 55 60 40

क्या ये ब्रांड औसत जीवन के संबंध में भिन्न हैं?

हम पहले दो नमूनों के नमूना प्रयोगाश्रित वितरण की गणना करते हैं:

x	$F_6(x)$	$G_6(x)$	$F_6(x) - G_6(x)$
30	2/6	0	2/6
40	4/6	1/6	3/6
45	5/6	2/6	3/6
50	5/6	4/6	1/6
55	1	5/6	1/6
60	1	1	0

$D_{m,n}$= sup $|F_6(x) - G_6(x) =|$ =3/6, तालिका से α = .05 के स्तर पर m = n = 6 के लिए महत्वपूर्ण मान है 4/6. चूँकि $D_{m,n}$ का परिकलित मान, n सारणीबद्ध मान से अधिक नहीं है, H_0

को अस्वीकार नहीं किया जाता है और यह निष्कर्ष निकाला जाता है कि दो ब्रांडों के लिए जीवन की औसत लंबाई समान है।

References:

Bhattacharya, G.K. and Johnson, R.A. *Statistics concepts and Methods.* New York, John Wiley and Sons. pp 505-521.

Neave, H.R. and Worthington, P.L. *Distribution free tests.* London Unwin Hyman, 161-164,328,337-341.

Neter, J.W.W. and Whitemore, G.A. *applied Statistics.* London, Allyn and Bacon Inc. 360- 388.

Ostle, B. *Statistics in Reasersch.* Ames. Iowa, USA. The Iowa State University.466-473.

सांख्यिकी में पायथन प्रोग्रामिंग का परिचय

प्रकाश कुमार
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
prakash.kumar@icar.gov.in

1. परिचय

सांख्यिकी गणित की वह शाखा है जो डेटा के संग्रह, संगठन, विश्लेषण, व्याख्या और प्रतिनिधित्व से संबंधित है। इस व्याख्यान को लेने के बाद, प्रतिभागी सांख्यिकीय विश्लेषण से संबंधित बुनियादी विश्लेषण कार्यों का वर्णन और प्रदर्शन करने में सक्षम होंगे। प्रतिभागी मौलिक सांख्यिकीय अवधारणाओं की समझ प्रदर्शित करने में सक्षम होंगे एवं आनुवंशिकी विश्लेषण में उपयोग किए जाने वाले बुनियादी आंकड़ों को समझने में सक्षम होंगे।
पायथन सामान्य प्रयोजन के अनुप्रयोगों के लिए एक लोकप्रिय उच्च स्तरीय प्रोग्रामिंग भाषा है। गुइडो वैन रोसुम ने 1991 में इसका आविष्कार किया और पायथन सॉफ्टवेयर फाउंडेशन ने इसे विकसित किया। पायथन एक प्रोग्रामिंग भाषा है जो आपको सिस्टम के साथ अधिक तेज़ी से और कुशलता से संचालित करने की अनुमति देती है। यह जटिल अनुप्रयोगों के त्वरित प्रोटोटाइप के लिए एकदम सही है। पायथन का उपयोग कई महत्वपूर्ण निगमों द्वारा किया जाता है, जिनमें नासा, गूगल, यूट्यूब, बिटटोरेंट और अन्य शामिल हैं। पायथन 2.7 और पायथन 3 दो सबसे लोकप्रिय पायथन संस्करण हैं। पायथन प्रोग्रामिंग को हमारे कार्यक्रमों की व्याख्या करने और चलाने के लिए एक इंटरप्रेटर की आवश्यकता होती है। पायथन स्क्रिप्ट चलाने के लिए स्वतंत्र रूप से कई इंटरप्रेटर उपलब्ध हैं जैसे IDLE (एकीकृत विकास पर्यावरण) जो तब इनस्टॉल होता है जब आप http://python.org/downloads/ से पायथन सॉफ्टवेयर इनस्टॉल करते हैं।

पायथन प्रोग्रामिंग विशेषताएँ:

- ✓ पायथन ऑब्जेक्ट ओरिएंटेड भाषा है: संरचना बहुरूपता, ऑपरेशन ओवरलोडिंग और मल्टीपल इनहेरिटेंस जैसी अवधारणाओं का समर्थन करती है।
- ✓ इंडेंटेशन: इंडेंटेशन पायथन में सबसे बड़ी विशेषताओं में से एक है
- ✓ यह मुफ़्त है (खुला स्रोत): पायथन को डाउनलोड करना और पायथन को इनस्टॉल करना मुफ़्त और आसान है
- ✓ लाइब्रेरी यूटिलिटीज: थर्ड पार्टी यूटिलिटीज (जैसे न्यूमेरिक(Numeric), नमपाय(NumPy), साइपाय(sciPy), स्कायकिट-लर्न(Scikit-learn), टेन्सरफ्लो(TensorFlow), केरास(Keras), पायटॉर्च(PyTorch), पांडा(Pandas) और मैटप्लोटलिब(Matplotlib) आदि)
- ✓ यह पोर्टेबल है: पायथन आज इस्तेमाल होने वाले लगभग हर बड़े प्लेटफॉर्म पर चलता है
- ✓ इसका उपयोग करना और सीखना आसान है
- ✓ इंटरप्रेटेड भाषा: पायथन को पायथन इंटरप्रेटर द्वारा रनटाइम पर संसाधित किया जाता है
- ✓ इंटरएक्टिव प्रोग्रामिंग भाषा: उपयोगकर्ता प्रोग्राम लिखने के लिए सीधे पायथन इंटरप्रेटर के साथ बातचीत कर सकते हैं

इंस्टालेशन(Installation):

http://python.org/downloads/ साइट से डाउनलोड कर पायथन सॉफ्टवेयर इनस्टॉल करते हैं।

विंडोज़ में पायथन इनस्टॉल करने के लिए अपनाए जाने वाले कदम:

चरण 1: पायथन करने के लिए पायथन के संस्करण का चयन करें।

चरण 2: पायथन निष्पादन योग्य इंस्टॉलर डाउनलोड करें।
चरण 3: निष्पादन योग्य इंस्टॉलर चलाएँ।
चरण 4: सत्यापित करें कि विंडोज पर पायथन पायथन हो गया।

इंटरैक्टिव मोड में पायथन चलाना:
पायथन स्क्रिप्ट फ़ाइल को इंटरप्रेटर को पास किए बिना, सीधे पायथन प्रॉम्प्ट पर कोड निष्पादित करें। एक बार जब आप पायथन इंटरप्रेटरके अंदर हों, तो आप शुरू कर सकते हैं।

```
>>> print("hello world")

hello world
```

वैकल्पिक रूप से, प्रोग्रामर पायथन स्क्रिप्ट स्रोत कोड को input.py फ़ाइल में सहेज सकते हैं और इंटरप्रेटर का उपयोग करके फ़ाइल की कोड को चला सकते हैं। स्क्रिप्ट चलाने के लिए आपको इंटरप्रेटर को फ़ाइल का नाम बताना होगा। उदाहरण के लिए, यदि आपके पास input.py नाम की एक स्क्रिप्ट है और आप यूनिक्स का उपयोग कर रहे हैं, तो आपको इसे चलाने के लिए python input.py टाइप करना होगा।

```
# To verify the type of any object in Python, use the type() function:
# पायथन में किसी भी ऑब्जेक्ट के प्रकार को सत्यापित करने के लिए, प्रकार (type()) फ़ंक्शन का उपयोग करें:
>>> type(10)

<class 'int'>

>>> a=11

>>> print(type(a))

<class 'int'>

>>> y=2.8

>>> print(type(y))

<class 'float'>

>>> type("hello world")

<class 'str'>

>>> list=[1,2,'A','B',[10,11]]

>>> type(list)

<class 'list'>

>>> tuple1=(1,2,3,4)

>>> type(tuple1)

<class 'tuple'>

>>> x=list(tuple1)

>>> x

[1, 2, 3, 4]
```

पायथन चर लिखने के नियम:

• एक चर नाम एक अक्षर या अंडरस्कोर वर्ण से शुरू होना चाहिए
• एक चर नाम एक संख्या से शुरू नहीं हो सकता
• एक चर नाम में केवल अल्फा-न्यूमेरिक वर्ण और अंडरस्कोर हो सकते हैं (A-z, 0-9,
तथा _)
• चर नाम केस-संवेदी होते हैं (age, Age and AGE तीन अलग-अलग चर हैं)

पायथन में प्रयुक्त ऑपरेटर

जोड़ें	+
घटाना	-
गुणा करें	*
पूर्णांक प्रभाग	/
शेष	%
बाइनरी लेफ्ट शिफ्ट	<<
बाइनरी राइट शिफ्ट	>>
तथा	&
या	\
से कम	<
से बड़ा	>
से कम या उसके बराबर	<=
से बड़ा या उसके बराबर	>=
समानता की जाँच करें	==
चेक बराबर नहीं	!=

टिप्पणी पंक्ति का प्रतीक (Comment line symbol):

एकल-पंक्ति टिप्पणियां हैश (#) प्रतीक से शुरू होती हैं और यह उल्लेख करने में उपयोगी होती हैं कि
पूरी लाइन को लाइन के अंत तक एक टिप्पणी के रूप में माना जाना चाहिए।
जब हमें कई पंक्तियों पर टिप्पणी करने की आवश्यकता होती है तो एक बहु-पंक्ति टिप्पणी उपयोगी होती है। पायथन में, ट्रिपल डबल कोट ("" ") और सिंगल कोट (' ') का उपयोग बहु-पंक्ति टिप्पणी के लिए किया जाता है।

पायथन मॉड्यूल (Python module):

पायथन मॉड्यूल को एक पायथन प्रोग्राम की गई फ़ाइल के रूप में वर्णित किया जाता है जिसमें पायथन कोड होता है, जैसे कि फ़ंक्शन, क्लासेस या चर। इसे दूसरे तरीके से रखने के लिए, एक्सटेंशन (.py) के साथ हमारी पायथन कोड फ़ाइल को मॉड्यूल माना जाता है। पायथन मॉड्यूल में निष्पादन योग्य कोड हो सकता है। पायथन में, एक मॉड्यूल हमें अपने कोड को तार्किक तरीके से संरचित करने की क्षमता देता है। एक मॉड्यूल की क्षमताओं का दूसरे में उपयोग करने के लिए हमें संबंधित मॉड्यूल को इम्पोर्ट करना होता है।

वाक्य - विन्यास:

```
#import <module-name>
```

```
>>> import sys
```

```
>>> print(sys.version)
```

```
3.9.5 (default, Nov 18 2021, 16:00:48)
```

[GCC 10.3.0]

फलन और उनके अनुप्रयोग(Functions and their applications):
फ़ंक्शन एक कनेक्टेड स्टेटमेंट का एक संग्रह है जो एक ही कार्य को निष्पादित करता है। फ़ंक्शंस हमारे सॉफ़्टवेयर को छोटे, मॉड्यूलर भागों में विभाजित करने में सहायता करते हैं। जैसे-जैसे यह आकार में बढ़ता है, फ़ंक्शंस हमारे प्रोग्राम को अधिक व्यवस्थित और नियंत्रित करने में मदद करते हैं। यह समय बचाता है और कोड को अधिक पुन: प्रयोज्य बनाता है।
मूल रूप से, हम फ़ंक्शन को निम्नलिखित दो प्रकारों में विभाजित कर सकते हैं:

1. बिल्ट-इन फंक्शन (Built-in functions) - फंक्शन जो पायथन में निर्मित होते हैं।

Ex: abs(), all(). Ascii(), bool().....so on....

integer = -20

print('Absolute value of -20 is:', abs(integer))

Output:

Absolute value of -20 is: 20

2. उपयोगकर्ता-परिभाषित फलन(User-defined functions) - स्वयं उपयोगकर्ताओं द्वारा परिभाषित फलन।

```
>>> def add_numbers(x,y):
...     sum = x + y
...     return sum
...
print("The sum is", add_numbers(5, 10))
```

आउटपुट होगा:
The sum is 15

Loops in Python:

A loop statement allows us to execute a statement or group of statements multiple times as long as the condition is true. Repeated execution of a set of statements with the help of loops is called iteration.

In Python Iteration (Loops) statements are of three types:

1. While Loop

2. For Loop

3. Nested For Loops

पायथन में लूप्स:
लूप स्टेटमेंट हमें स्टेटमेंट या स्टेटमेंट के समूह को कई बार निष्पादित करने की अनुमति देता है, जब तक कि कंडीशन सही है। लूप की मदद से बयानों के एक सेट के बार-बार निष्पादन किया जाता है।
पायथन इटरेशन (लूप्स) में कथन तीन प्रकार के होते हैं:
1. जबकि लूप (While Loop)
2. फॉर लूप (For Loop)
3. नेस्टेड फॉर लूप्स (Nested For Loops)

```
>>> i=1
```

```
>>> while i<=5:
...     print(i)
...     i=i+1
...
1
2
3
4
5
>>> numbers=[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]
>>> for i in numbers:
...     square=i*i
...     print(square)
...
1
4
16
36
121
400
>>> for i in range(len(numbers)):
...     square=i*i
...     print(square)
```

सूची ऑब्जेक्ट के तरीके (Methods of list objects):

Del(), Apend(), Extend(), Insert(), Pop(), Remove(), Reverse() and Sort()

हटाना (Delete):

Del()

```
>>> x=[5,3,8,6]
>>> del(x[1]) #deletes the index position 1 in a list
>>> x
[5, 8, 6]
```

संलग्न (Append)

```
>>> x=[1,5,8,4]
>>> x.append(10)
>>> x
[1, 5, 8, 4, 10]
```

बढ़ाएँ (Extend): किसी सूची में अनुक्रम जोड़ें।

```
>>> x=[1,2,3,4]
>>> y=[3,6,9,1]
>>> x.extend(y)
>>> x
[1, 2, 3, 4, 3, 6, 9, 1]
```

सम्मिलित करें (Insert): निर्दिष्ट अनुक्रमणिका में कोई आइटम जोड़ने के लिए, insert () विधि का उपयोग करें:

```
>>> x=[1,2,4,6,7]
>>> x.insert(2,10) #insert(index no, item to be inserted)
>>> x
[1, 2, 10, 4, 6, 7]
>>> x.insert(4,['a',11])
>>> x
[1, 2, 10, 4, ['a', 11], 6, 7]
```

पॉप (Pop): pop() विधि निर्दिष्ट अनुक्रमणिका को हटा देती है, (या अंतिम आइटम यदि अनुक्रमणिका नहीं है निर्दिष्ट) या बस सूची के अंतिम आइटम को पॉप करता है और आइटम लौटाता है।

```
>>> x=[1, 2, 10, 4, 6, 7]
>>> x.pop()
7
>>> x
[1, 2, 10, 4, 6]
>>> x=[1, 2, 10, 4, 6]
>>> x.pop(2)
10
>>> x
[1, 2, 4, 6]
```

निकालें (remove): remove() विधि निर्दिष्ट सूची से निर्दिष्ट आइटम को हटा देती है।

```
>>> x=[1,33,2,10,4,6]
>>> x.remove(33)
>>> x
```

[1, 2, 10, 4, 6]

रिवर्स (reverse): दी गई सूची के क्रम को उलट दें।

```
>>> x=[1,2,3,4,5,6,7]

>>> x.reverse()

>>> x

[7, 6, 5, 4, 3, 2, 1]
```

क्रमबद्ध करें (sort): सूची के तत्वों को आरोही क्रम में क्रमबद्ध करें

```
>>> x=[7, 6, 5, 4, 3, 2, 1]

>>> x.sort()

>>> x

[1, 2, 3, 4, 5, 6, 7]
```

डिक्शनरी (Dictionaries):

डिक्शनरी एक संग्रह है जो अनियंत्रित, परिवर्तनशील और अनुक्रमित है। पायथन शब्दकोशों में करली ब्रैकेट के साथ लिखा जाता है, और उनके पास कुंजियाँ और मान होते हैं।

- कुंजी- मान जोड़े (Key-value pairs)
- अव्यवस्थित (Unordered)

```
>>> dict1 = {"a":1,"b":"college","year":2004}

>>> dict1

{'a': 1, 'b': 'college', 'year': 2004}

>>> for k,v in dict1.items():

...     print(k,v)

...

a 1

b college

year 2004
```

पायथन में बुनियादी सांख्यिकी (Basic Statistics in Python):

पायथन की सहायता से माध्य, माध्यिका, बहुलक और मानक विचलन जैसे आँकड़ों की मूल बातें यहाँ वर्णित की गई हैं। उदाहरण के लिए

माध्य (Mean):

यहाँ माध्य संख्याओं के औसत को दर्शाता है, जिसका अर्थ है कि हम संख्याओं को जोड़ते हैं और उन्हें उपस्थित वस्तुओं की कुल संख्या से विभाजित करते हैं। इसके लिए कोड है:

```
a=[11, 21, 34, 22, 27, 11, 23, 21]
mean = sum(a)/len(a)

print (mean)

# numpy python modue का उपयोग करके
```

```
import numpy as np
a =[11, 21, 34, 22, 27, 11, 23, 21]
mean = np.mean(a)
print (mean)
```

माध्यिका (Median):

माध्यिका मध्य पद है जो एक क्रमबद्ध सरणी में होता है। एक सूची के तत्वों की एक विषम संख्या के लिए, माध्यिका मध्य पद है, और एक सूची के तत्वों की एक सम संख्या के लिए, माध्यिका मध्य में दो पदों का औसत होता है।

```
def median(nums):
    nums.sort()
    if len(nums)%2 == 0:
        return int(nums[len(nums)//2-1]+nums[len(nums)//2])/2
    else:
        return nums[len(nums)//2]

a =[11, 21, 34, 22, 27, 11, 23, 21]
print (median(a))
# या numpy module का उपयोग से
import numpy as np
a =[11, 21, 34, 22, 27, 11, 23, 21]
print(np.median(a))
```

चतुर्थक (Quartiles):

चतुर्थक आँकड़ों को चार भागों में विभाजित करते हैं। पहले भाग में प्रारंभ से प्रथम चतुर्थक (Q1), दूसरे भाग में प्रथम चतुर्थक से द्वितीय चतुर्थक (Q2), तीसरा भाग Q2 से Q3 और चौथा भाग Q3 से अंत तक शामिल है। चतुर्थक खोजने के लिए डेटा को क्रमबद्ध किया जाना चाहिए।

```
def quartiles(nums):
    nums=sorted(nums)
    Q1 = median(nums[:len(nums)//2])
    Q2 = median(nums)
    if len(nums)%2 == 0:
        Q3 = median(nums[len(nums)//2:])
    else:
        Q3 = median(nums[len(nums)//2+1:])
```

```
    return Q1,Q2,Q3


def median(nums):
    nums.sort()
    if len(nums)%2 == 0:


        return int(nums[len(nums)//2-1]+nums[len(nums)//2])/2
    else:
        return nums[len(nums)//2]


a =[11, 21, 34, 22, 27, 11, 23, 21]
print (quartiles(a))
```

मानक विचलन (Standard deviation):

मानक विचलन डेटा के फैलाव या प्रसार का माप है। यह वेरिएंस का वर्गमूल है। मानक विचलन निकलने के लिए सरल पायथन कोड नीचे दिया गया है:

```
import numpy as np
a =[11, 21, 34, 22, 27, 11, 23, 21]
print (np.std(a))
# पायथन मॉड्यूल और मानक  लाइब्रेरी का उपयोग करके
import numpy as np
from scipy import stats


import matplotlib
import matplotlib.pyplot as plt
matplotlib.style.use('ggplot')


np.random.seed(1)
data = np.round(np.random.normal(5, 2, 100))
plt.hist(data, bins=10, range=(0,10), edgecolor='black')
plt.show()
mean = np.mean(data)
mean
np.median(data) # for median
```

```
mode = stats.mode(data) #for mode

print("The modal value is {} with a count of {}".format(mode.mode[0], mode.count[0]))

np.ptp(data) # range

np.var(data) # for varience

np.std(data) # for standard deviations etc.
```

डेटा विश्लेषण के लिए महत्वपूर्ण पायथन लाइब्रेरी

आपके डेटा विश्लेषण के लिए उपयोग की जाने वाली प्रमुख पायथन लाइब्रेरी नीचे दी गई हैं।

Numpy- मौलिक वैज्ञानिक कंप्यूटिंग

NumPy, वैज्ञानिक कंप्यूटिंग के लिए सबसे महत्वपूर्ण पायथन मॉड्यूल है। यह एक पायथन लाइब्रेरी है जिसमें एक बहुआयामी ऐरे ऑब्जेक्ट, व्युत्पन्न ऑब्जेक्ट (जैसे मैट्रिसेस), और गणितीय, तार्किक, आकार में हेरफेर, छँटाई, चयन, I/O, जैसे तेज़ ऐरे संचालन करने के लिए विभिन्न प्रकार के मॉड्यूल जैसे की असतत फूरियर रूपांतरण, बुनियादी रैखिक बीजगणित, बुनियादी सांख्यिकीय संचालन, यादृच्छिक अनुकरण, और बहुत कुछ शामिल हैं।
SciPy का मतलब साइंटिफिक पायथन है। यह NumPy पर बनाया गया है। असतत फूरियर रूपांतरण, रैखिक बीजगणित, अनुकूलन और विरल मैट्रिक्स जैसे उच्च स्तरीय विज्ञान और इंजीनियरिंग मॉड्यूल की विविधता के लिए Scipy सबसे उपयोगी लाइब्रेरी में से एक है।
उदाहरण:

```
>>> import numpy as np

>>> a = np.arange(15).reshape(3, 5)

>>> a

array([[ 0,  1,  2,  3,  4],

       [ 5,  6,  7,  8,  9],

       [10, 11, 12, 13, 14]])

>>> a.shape

(3, 5)

>>> a.ndim

2

>>> a.dtype.name

'int64'

>>> type(a)

<class 'numpy.ndarray'>

>>> b = np.array([6, 7, 8])

>>> b

array([6, 7, 8])

>>> b = np.array([1.2, 3.5, 5.1])
```

```
>>> b.dtype

dtype('float64')

>>> mat = np.matrix('1 2; 3 4')

>>> mat.T

matrix([[1, 3],

        [2, 4]])
```

Scipy - मौलिक वैज्ञानिक कम्प्यूटिंग

NumPy, जो सुविधाजनक और त्वरित N-आयामी सरणी हेरफेर को सक्षम बनाता है, SciPy पैकेज द्वारा उपयोग किया जाता है। SciPy लाइब्रेरी को NumPy सरणियों के साथ संचालित करने के लिए डिज़ाइन किया गया था और इसमें कई उपयोगकर्ता के अनुकूल और कुशल संख्यात्मक प्रक्रियाएं शामिल हैं, जैसे संख्यात्मक एकीकरण और अनुकूलन विधियां। वे सभी प्रमुख ऑपरेटिंग सिस्टम पर काम करते हैं, इनस्टॉल करने में आसान हैं, और पूरी तरह से निःशुल्क हैं।

SciPy को विभिन्न वैज्ञानिक कंप्यूटिंग डोमेन को कवर करने वाले उप-पैकेजों में व्यवस्थित किया गया है। इन्हें नीचे दिए गए में संक्षेपित किया गया है

- वेक्टर परिमाणीकरण के लिए scipy.cluster / Kmeans
- भौतिक और गणितीय स्थिरांक के लिए scipy.constants
- फूरियर रूपांतरण के लिए scipy.fftpack
- एकीकरण दिनचर्या के लिए scipy.integrate
- इंटरपोलेशन के लिए scipy.interpolate
- डेटा इनपुट और आउटपुट के लिए scipy.io
- रैखिक बीजगणित दिनचर्या के लिए scipy.linalg
- n-आयामी छवि पैकेज के लिए scipy.ndimage
- ऑर्थोगोनल दूरी प्रतिगमन के लिए scipy.odr
- अनुकूलन के लिए scipy.optimize
- सिग्नल प्रोसेसिंग के लिए scipy.signal
- विरल मैट्रिसेस के लिए scipy.sparse
- स्थानिक डेटा संरचनाओं और एल्गोरिदम के लिए scipy.spatial
- किसी विशेष गणितीय कार्यों के लिए scipy.special
- सांख्यिकी के लिए scipy.stats

उदाहरण:

```
>>> from scipy.constants import pi

>>> from math import pi

>>> print("pi = %.16f"%scipy.constants.pi)

pi = 3.1415926535897931

#Interpolation is the process of finding a value between two points on a line or a curve.

>>> import numpy as np

>>> from scipy import interpolate

>>> import matplotlib.pyplot as plt

>>> x = np.linspace(0, 4, 12)

>>> y = np.cos(x**2/3+4)

>>> print(x,y)
```

```
(
  array([0.,  0.36363636,  0.72727273,  1.09090909,  1.45454545, 1.81818182,
         2.18181818,  2.54545455,  2.90909091,  3.27272727,  3.63636364,  4.]),

  array([-0.65364362,  -0.61966189,  -0.51077021,  -0.31047698,  -0.00715476,
          0.37976236,   0.76715099,   0.99239518,   0.85886263,   0.27994201,
         -0.52586509,  -0.99582185])
)
```

पांडा (Pandas) - डेटा मेनूपुलेशन और विश्लेषण हेतु :

संरचित डेटा संचालन और जोड़तोड़ के लिए Pandas लाइब्रेरी का उपयोग किया जाता है। Numpy, Scipy, Cython, और Pandas त्वरित डेटा प्रोसेसिंग टूल उपलब्ध हैं। हालांकि, हम Pandas के पक्ष में हैं क्योंकि वे अन्य टूल की तुलना में तेज़, आसान और अधिक अभिव्यंजक हैं। Pandas को अपेक्षाकृत हाल ही में पायथन में जोड़ा गया था और डेटा वैज्ञानिक समुदाय में पायथन के उपयोग को बढ़ाने में महत्वपूर्ण भूमिका निभाई है।

उदाहरण:

```
>>> import pandas as pd
>>> import numpy as np
>>> info = np.array(['P','a','n','d','a','s'])
>>> a = pd.Series(info)
>>> print(a)
0   P
1   a
2   n
3   d
4   a
5   s
dtype: object
>>> data = [['Alex',10],['Bob',12],['Clarke',13]]
>>> df = pd.DataFrame(data,columns=['Name','Age'])
>>> print (df)
    Name  Age
0   Alex   10
1    Bob   12
2 Clarke   13
>>> data = {'Name':['Tom', 'Jack', 'Steve', 'Ricky'],'Age':[28,34,29,42]}
```

```
>>> df = pd.DataFrame(data)
>>> print (df)
   Name  Age
0   Tom   28
1  Jack   34
2 Steve   29
3 Ricky   42
>>> data = [{'a': 1, 'b': 2},{'a': 5, 'b': 10, 'c': 20}]
>>> df = pd.DataFrame(data)
>>> print (df)
  a   b     c
0 1   2   NaN
1 5  10  20.0
```

Make interactive figures that can zoom, pan, update.

- ✓ Customize visual style and layout.
- ✓ Export to many file formats .
- ✓ Embed in JupyterLab and Graphical User Interfaces.
- ✓ Use a rich array of third-party packages built on Matplotlib.

Matplotlib - प्लॉटिंग और विज़ुअलाइज़ेशन के लिए :

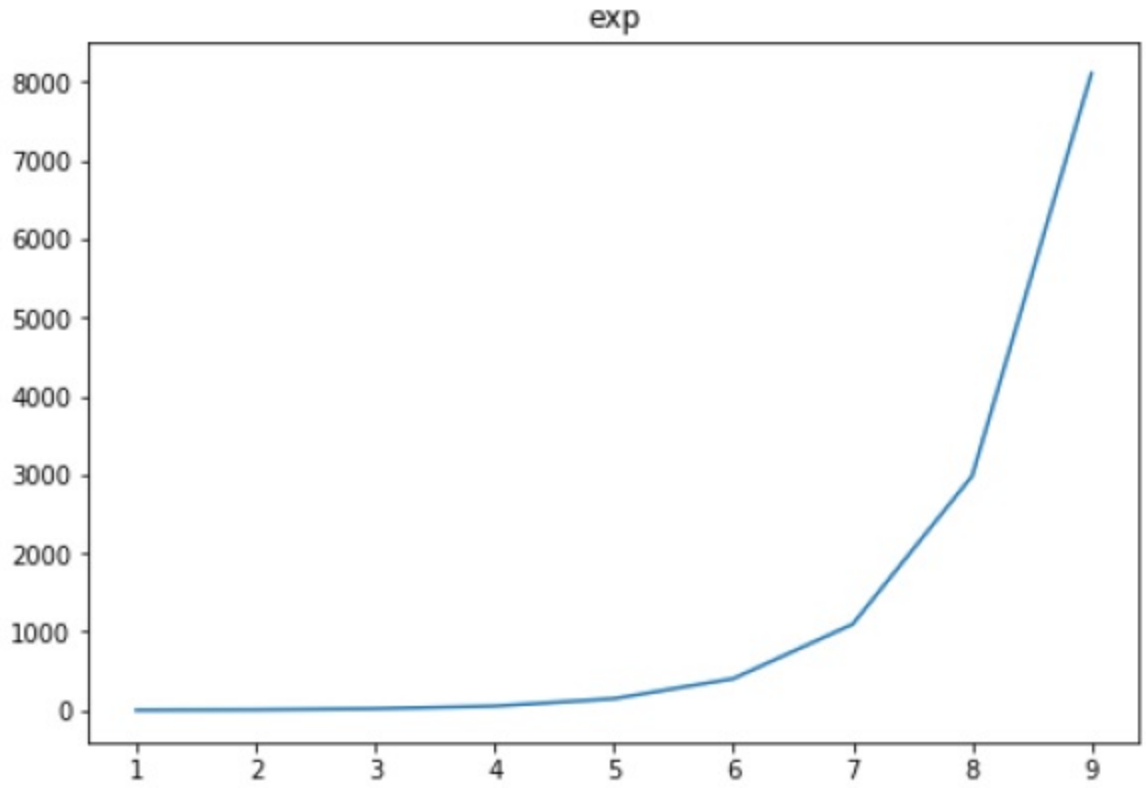
हिस्टोग्राम, लाइन प्लॉट से लेकर हीट मैप प्लॉट तक, विभिन्न प्रकार के प्लॉटिंग करने के लिए Matplotlib का उपयोग करते है । Matplotlib पायथन में स्थिर, एनिमेटेड और इंटरैक्टिव विज़ुअलाइज़ेशन के लिए एक व्यापक लाइब्रेरी है। Matplotlib कठिन चीजों को आसान बनाता है।

- ✓ प्रकाशन गुणवत्ता वाले प्लॉट बना सकते है ।
- ✓ इंटरैक्टिव प्लाट बनाएं जिसे ज़ूम, पैन, अपडेट कर सकते हैं।
- ✓ दृश्य शैली और लेआउट को अनुकूलित कर सकते हैं।
- ✓ कई फ़ाइल स्वरूपों में एक्सपोर्ट कर सकते हैं।
- ✓ JupyterLab और ग्राफिकल यूजर इंटरफेस में एम्बेड कर सकते हैं।

उदाहरण :

```
>>> import matplotlib.pyplot as plt
>>> fig = plt.figure()
>>> a1 = fig.add_axes([0,0,1,1])
>>> import numpy as np
>>> x = np.arange(1,10)
```

```
>>> a1.plot(x, np.exp(x))
[<matplotlib.lines.Line2D object at 0x7fc3f91a9910>]
>>> a1.set_title('exp')
Text(0.5, 1.0, 'exp')
>>> plt.show()
```

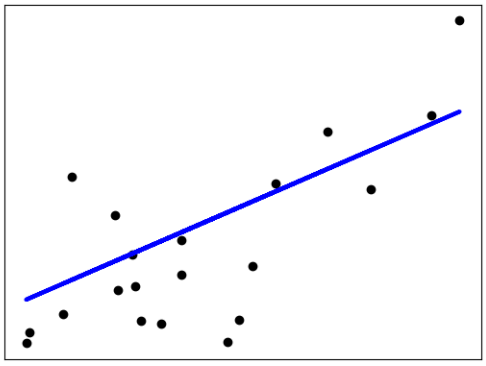


साईकिट-लर्न (Scikit-learn)- मशीन लर्निंग और डेटा माइनिंग के लिए

मशीन लर्निंग के लिए साईकिट लर्न, NumPy, SciPy और matplotlib पर निर्मित है, इस लाइब्रेरी में वर्गीकरण, प्रतिगमन, क्लस्टरिंग और आयामी कमी सहित मशीन लर्निंग और सांख्यिकीय मॉडलिंग के लिए बहुत सारे कुशल उपकरण हैं।
उदाहरण, रैखिक प्रतिगमन:

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> from sklearn import datasets, linear_model
>>> from sklearn.metrics import mean_squared_error, r2_score
>>> diabetes_X, diabetes_y = datasets.load_diabetes(return_X_y=True)
>>> diabetes_X = diabetes_X[:, np.newaxis, 2]
>>> # Split the data into training/testing sets
>>> diabetes_X_train = diabetes_X[:-20]
```

```
>>> diabetes_X_test = diabetes_X[-20:]
>>> diabetes_y_train = diabetes_y[:-20]
>>> diabetes_y_test = diabetes_y[-20:]
# Create linear regression object
>>> regr = linear_model.LinearRegression()
>>> # Train the model using the training sets
>>> regr.fit(diabetes_X_train, diabetes_y_train)
LinearRegression()
>>> diabetes_y_pred = regr.predict(diabetes_X_test)
>>> print("Coefficients: \n", regr.coef_)
Coefficients:
 [938.23786125]
# The mean squared error
>>> print("Mean squared error: %.2f" % mean_squared_error(diabetes_y_test, diabetes_y_pred))
Mean squared error: 2548.07
>>> # The coefficient of determination: 1 is perfect prediction
>>> print("Coefficient of determination: %.2f" % r2_score(diabetes_y_test, diabetes_y_pred))
Coefficient of determination: 0.47
>>> plt.scatter(diabetes_X_test, diabetes_y_test, color="black")
<matplotlib.collections.PathCollection object at 0x7fc3c70c2460>
>>> plt.plot(diabetes_X_test, diabetes_y_pred, color="blue", linewidth=3)
[<matplotlib.lines.Line2D object at 0x7fc3c70c2c40>]
>>> plt.xticks(())
([], [])
>>> plt.yticks(())
([], [])
>>>
>>> plt.show()
```



एसवीएम का उदाहरण (Example of SVM):

```
>>> import numpy as np
>>> from sklearn.datasets import make_classification
>>> from sklearn.model_selection import train_test_split
>>> from sklearn import svm
>>> X, y = make_classification(n_samples=10, random_state=0)
>>> X_train , X_test , y_train, y_test = train_test_split(X, y, random_state=0)
>>> clf = svm.SVC(kernel='precomputed')
>>> gram_train = np.dot(X_train, X_train.T)
>>> clf.fit(gram_train, y_train)
SVC(kernel='precomputed')
>>> # predict on training examples
>>> gram_test = np.dot(X_test, X_train.T)
>>> clf.predict(gram_test)
array([0, 1, 0])
```

StatsModels - सांख्यिकीय मॉडलिंग, परीक्षण और विश्लेषण

सांख्यिकीय मॉडलिंग के लिए Statsmodels। यह एक पायथन मॉड्यूल है जो उपयोगकर्ताओं को डेटा का पता लगाने, सांख्यिकीय मॉडल का अनुमान लगाने और सांख्यिकीय परीक्षण करने की अनुमति देता है। विभिन्न प्रकार के डेटा और प्रत्येक अनुमानक के लिए वर्णनात्मक सांख्यिकी, सांख्यिकीय परीक्षण और प्लॉटिंग फ़ंक्शन की एक विस्तृत सूची उपलब्ध है।

सीबॉर्न (Seaborn) - सांख्यिकीय डेटा विज़ुअलाइज़ेशन के लिए

सांख्यिकीय डेटा विज़ुअलाइज़ेशन के लिए सीबॉर्न उपयोगी है । यह पायथन में आकर्षक और सूचनात्मक सांख्यिकीय ग्राफिक्स बनाने के लिए एक लाइब्रेरी है। यह matplotlib पर आधारित है। सीबॉर्न का उद्देश्य विज़ुअलाइज़ेशन को डेटा की खोज और समझ का एक केंद्रीय हिस्सा बनाना है।

संदर्भ:

TEXT BOOKS

Allen B. Downey, ``Think Python: How to Think Like a Computer Scientist'', 2nd edition, Updated for Python 3, Shroff/O'Reilly Publishers, 2016.

R. Nageswara Rao, "Core Python Programming", dreamtech

Python Programming: A Modern Approach, Vamsi Kurama, Pearson

Core Python Programming, W.Chun, Pearson.

Introduction to Python, Kenneth A. Lambert, Cengage

Learning Python, Mark Lutz, Orielly

The Python Tutorial, **https://docs.python.org/3/tutorial/**

The Python Language Reference, **http://docs.python.org/3/reference/**

The Python Standard Library, **http://docs.python.org/3/library/**

PEP-8: Style Guide for Python Code, **http://www.python.org/dev/peps/pep-0008/**

Website: https://www.python.org

सांख्यिकीय आनुवंशिकी में मशीन लर्निंग तकनीकों का उपयोग

प्रबीन कुमार मेहर
भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012
prabina.Meher@icar.gov.in

मशीन लर्निंग आर्टिफिशियल इंटेलिजेंस की एक शाखा है। एमएल में सिस्टम को डेटा से जटिल पैटर्न सीखने के लिए कंप्यूटर–आधारित एल्गोरिदम विकसित किए जाते हैं और सीखे गए पैटर्न के आधार पर नए विशिष्ट के लिए पूर्वानुमान लगाया जाता हैं। पूर्वानुमान को दो वर्गों में वर्गीकृत किया जा सकता है।

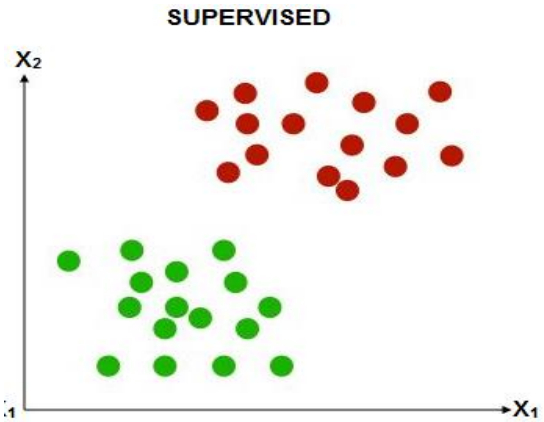
1 प्रेक्षणों के लेबल की पूर्वानुमान करना (बीमारी, कोई बीमारी नहीं)

2 मूल्यों की पूर्वानुमान करना, निरंतर या असतत (यील्ड)

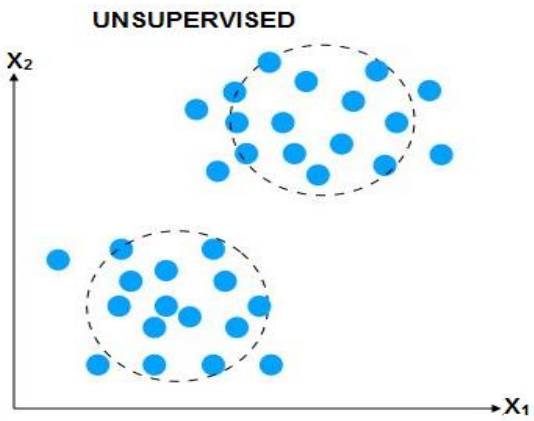
पहले प्रकार की पूर्वानुमान को वर्गीकृत वर्गीकरण और दूसरे प्रकार को प्रतिगमन के रूप में जाना जाता है। इसके अलावा, एमएल को मोटे तौर पर दो वर्गों में वर्गीकृत किया जा सकता है, अर्थात् पर्यवेक्षित शिक्षण और अनुपयोगी शिक्षण। ये दो सीखने की तकनीक मुख्य रूप से इनपुट–आउटपुट संबंध के आधार पर भिन्न हैं। दूसरे शब्दों में, पर्यवेक्षित शिक्षण एल्गोरिथम में लेबल (आउटपुट) प्रत्येक अवलोकन स जुड़े होते हैं जिसका हम अनुमान लगाना चाहते हैं (वर्गीकरण या प्रतिगमन), जबकि अनुपयोगी सीखने के मामले में अवलोकनों से कोई लेबल नहीं जुड़ा होता है और यहाँ हमारा उद्देश्य है मुख्य रूप से अवलोकनों का समूह है।

	Predictors			
Target (Y)	X_1	X_2	...	X_p
y_1	x_{11}	x_{12}	...	x_{1p}
y_2	x_{21}	x_{22}	...	x_{2p}
y_3	x_{31}	x_{32}	...	x_{3p}
...	...	...	...	...
y_n	x_{n1}	x_{n2}	x_{n3}	x_{np}

	Predictors			
	X_1	X_2	...	X_p
	x_{11}	x_{12}	...	x_{1p}
	x_{21}	x_{22}	...	x_{2p}
	x_{31}	x_{32}	...	x_{3p}
	...	...	...	...
	x_{n1}	x_{n2}	x_{n3}	x_{np}



Classification



Clustering

स्रोतः चित्र गूगल से लिए गए हैं

इस लेक्चर नोट में, हमारा ध्यान केवल पर्यवेक्षित मशीन लर्निंग एल्गोरिदम पर है। यहां, हम दो सामान्य रूप से उपयोग किए जाने वाले पर्यवेक्षित मशीन लर्निंग एल्गोरिदम पर चर्चा करेंगे जो सपोर्ट वेक्टर मशीन (एसवीएम) और रैंडम फॉरेस्ट (आरएफ) हैं।

माप की सटीकता

वर्गीकरण और प्रतिगमन समस्या में सटीकता को मापने के लिए विभिन्न प्रकार के मेट्रिक्स का उपयोग किया जाता है।

वर्गीकरण की सटीकता मेट्रिक्स

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+TN+FP}$$

$$\text{Sensitivity}=\frac{TP}{TP+FN}$$

$$\text{Specificity}=\frac{TN}{TN+FP}$$

$$\text{Precision}=\frac{TP}{TP+FP}$$

$$\text{F1-score}=2\times\frac{Precision\times recall}{Precision+Recall}$$

TP: true positive
TN: true negative
FP: false positive
FN: false negative

रिग्रेशन सटीकता मेट्रिक्स

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}|y-\hat{y}|$$

$$\text{Mean Square Error (MSE)} = \frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})^2$$

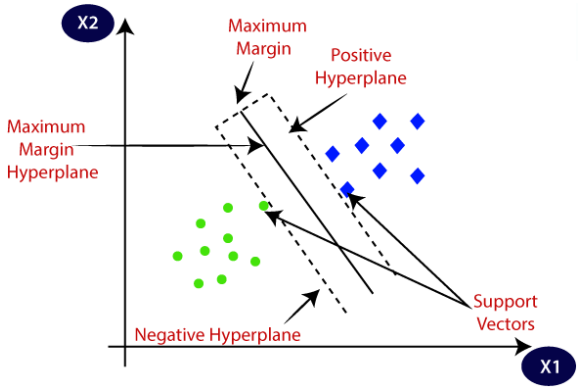
$$\text{Mean Percentage Error (MPE)} = \frac{1}{n}\sum_{i=1}^{n}\frac{(y-\hat{y})}{y}\times 100$$

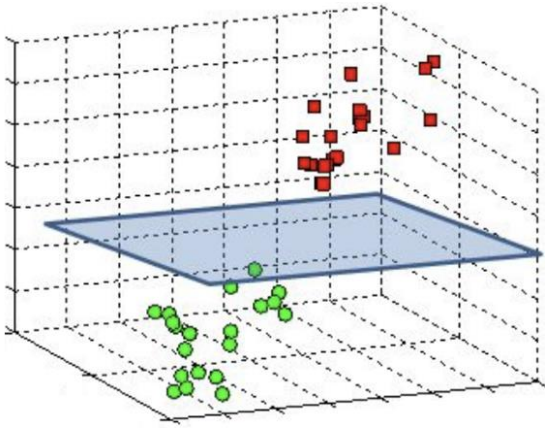
$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y-\hat{y}}{y}\right|\times 100$$

निम्नलिखित उपखंडों में, हम एसवीएम, आरएफ और एएनएन को लागू करने के लिए आर–कोड पर चर्चा करेंगे। इसके लिए यूजर को https://cran.r-project.org/bin/windows/base/. करना होगा।

समर्थन वेक्टर मशीन

सपोर्ट वेक्टर मशीन (एसवीएम) एक पर्यवेक्षित मशीन लर्निंग एल्गोरिदम है जिसका उपयोग वर्गीकरण और प्रतिगमन दोनों के लिए किया जा सकता है। वर्गीकरण के मामले में, एसवीएम की पूर्वानुमान कहनेवाला क्षमता काफी हद तक उपयोग किए जाने वाले कर्नेल फ़ंक्शन के प्रकार पर निर्भर करती है जो इनपुट डेटा को एक उच्च–आयामी सुविधा स्थान पर सेट करती है, जहां अवलोकन विभिन्न वर्गों से संबंधित होते हैं, जो इष्टतम पृथक्करण हाइपर–प्लेन द्वारा रैखिक रूप से अलग होते हैं। प्रतिगमन के मामले में, फिट की सबसे अच्छी रेखा की खोज की जाती है और यह और कुछ नहीं बल्कि अधिकतम अंक वाले हाइपर प्लेन है।





स्रोतः चित्र गूगल छवियों से लिए गए हैं

समर्थन वेक्टर मशीन वर्गीकरण आर में लागू करना

एसवीएम को लागू करने के लिए आर- पैकेज "e1071" इन्सटाल करें।

```
install.packages(caTools)
install.packages("e1071")
library(e1071)
library(caTools)
```

निर्देशिका से डेटा पढ़ें, जहां डेटा सहेजा गया है। यहां, हम आर यानी आईरिस डेटासेट में उपलब्ध इनबिल्ट डेटासेट का उपयोग करेंगे। यह एक बेंचमार्क डेटासेट है जिसमें तीन अलग–अलग फूलों की प्रजातियों के 150 अवलोकन शामिल हैं, प्रत्येक प्रकार के लिए 50 अवलोकनों के साथ ई सेटोसा, वर्सीकलर और वर्जिनिका। चार चर (पूर्वानुमान) हैं जैसे सीपल लंबाई, सेपल चौड़ाई, पंखुड़ी लंबाई और पंखुड़ी चौड़ाई। यह डेटा केवल आईरिस टाइप करके आर कंसोल में लोड किया जा सकता है।

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
```

चूंकि प्रतिक्रिया फूलों की विभिन्न प्रजातियां (लेबल) है, यह एक वर्गीकरण समस्या है। वर्गीकरण या प्रतिगमन के लिए, प्रशिक्षण और परीक्षण डेटासेट की आवश्यकता होती है। प्रशिक्षण सेट में संबंधित लेबल के साथ अवलोकन होते हैं जबकि परीक्षण सेट के अवलोकन के लिए लेबल की पूर्वानुमान की जाती है। इसलिए, पहले हमें डेटासेट को प्रशिक्षण और परीक्षण सेट में विभाजित करने की आवश्यकता है। प्रशिक्षण और परीक्षण के लिए उपयोग किए जाने वाले डेटासेट का प्रतिशत उपयोगकर्ता पर निर्भर करता है। हालांकि, मॉडल की बेहतर फिटिंग के लिए प्रशिक्षण डेटासेट काफी बड़ा होना चाहिए। इस उद्देश्य के लिए निम्न आदेश का उपयोग किया जा सकता है।

```
library(caTools)
set.seed(123)
dataset <- iris
part <- sample.split(dataset$Species, SplitRatio = 0.70)
train_set <- subset(dataset, part == TRUE)
test_set <- subset(dataset, part == FALSE)

#Check the number of observations of training and test set
table(train_set$Species)
table(test_set$Species)
```

प्रशिक्षण डेटासेट का उपयोग करके समर्थन वेक्टर मशीन वर्गीकरण मॉडल की फिटिंग

```
# Fitting SVM model to the Training set

library(e1071)
set.seed(123)
svm_class <- svm(formula = Species ~ .,
                 data = train_set,
                 type = 'C-classification',
                 kernel = 'radial')
```

मॉडल को प्रिंट करके कोई भी विवरण देख सकता है।

```
> svm_class

Call:
svm(formula = Species ~ ., data = train_set, type = "C-classification",
    kernel = "radial")


Parameters:
   SVM-Type:  C-classification
 SVM-Kernel:  radial
       cost:  1

Number of Support Vectors:  40
```

मॉडल फिटिंग में, अन्य मापदंडों के मूल्यों को डिफ़ॉल्ट रखा जाता है। हालांकि, उपयोगकर्ता को वर्गीकरण सटीकता को अधिकतम करने के लिए कर्नेल फ़ंक्शन के संबंधित पैरामीटर को ट्यून करना चाहिए। यहां, हमने रेडियल बेस कर्नेल (आरबीएफ) कर्नेल फ़ंक्शन का उपयोग किया है, लेकिन अन्य कर्नेल के साथ–साथ बहुपद, रैखिक और सिग्मॉइड का चयन करने का विकल्प है। वर्गीकरण सटीकता कर्नेल कार्यों के अनुसार भिन्न होती है। सपोर्ट वैक्टर वे अवलोकन हैं जो हाइपर–प्लेन के करीब होते हैं और साथ ही हाइपर–प्लेन की स्थिति और दिशा को प्रभावित करते हैं। इसलिए, समर्थन वैक्टर की संख्या हमेशा प्रशिक्षण टिप्पणियों की कुल संख्या से कम होती है। हाइपर–प्लेन के करीब से दिखने वाले प्रेक्षणों की अधिकतम संख्या मॉडल की बेहतर फिटिंग का प्रतिनिधित्व करती है। मॉडल को फिट करने के बाद, अगला भाग प्रशिक्षित मॉडल का उपयोग करके परीक्षण सेट (वर्गीकरण के लिए) के लेबल की पूर्वानुमान करना है। आइए परीक्षण सेट को प्रिंट और जांचें

```
> test_set
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
2            4.9         3.0          1.4         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
8            5.0         3.4          1.5         0.2     setosa
11           5.4         3.7          1.5         0.2     setosa
16           5.7         4.4          1.5         0.4     setosa
```

तो, कोई देख सकता है कि परोक्षण डेटासेट में लेबल (प्रजातियां) हैं। लेकिन, वास्तव में, लेबल नहीं होंगे और हमें पूर्वानुमान करने की आवश्यकता है। इस प्रकार, आइए परीक्षण डेटासेट से लेबल हटा दें

```
test_set1 <- test_set[-5]
```

```
> test_set
    Sepal.Length Sepal.Width Petal.Length Petal.Width
2            4.9         3.0          1.4         0.2
4            4.6         3.1          1.5         0.2
5            5.0         3.6          1.4         0.2
8            5.0         3.4          1.5         0.2
11           5.4         3.7          1.5         0.2
```

```
#Predicting the Test set results
y_pred = predict(svm_class, newdata = test_set1)
y_pred
```

```
> y_pred
         2          4          5          8         11         16
    setosa     setosa     setosa     setosa     setosa     setosa
        21         24         26         31         32         34
    setosa     setosa     setosa     setosa     setosa     setosa
        50         53         58         59         65         67
    setosa versicolor versicolor versicolor versicolor versicolor
```

यहां, हमने परीक्षण सेट के लेबल को पूर्वानुमान की है। कोई उस संभावना का अनुमान लगा सकता है जिसके साथ इन लेबलों की पूर्वानुमान की जाती है और इसके लिए मॉडल को प्रशिक्षित करते समय एक अलग तर्क पारित करने की आवश्यकता होती है।

```
#Training of the model with probability option
svm_class <- svm(formula = Species ~ .,
                 data = train_set,
                 type = 'C-classification',
                 kernel = 'radial', probability=TRUE)

#Prediction fro the test set with probability option
pred_prob <- predict(svm_class, newdata = test_set1, probability=TRUE)
```

```
> pred_prob
        setosa  versicolor   virginica
2   0.95757442 0.027914287 0.014511291
4   0.96313332 0.022506155 0.014360530
5   0.97177768 0.015992024 0.012230293
8   0.96951296 0.017704090 0.012782952
11  0.96913348 0.018803440 0.012063083
16  0.87234295 0.072760478 0.054896568
```

अगला कदम भ्रम मैट्रिक्स की गणना करना है जिसमें सही ढंग से वर्गीकृत और गलत वर्गीकृत टिप्पणियों की संख्या शामिल है। सबसे पहले, हम परीक्षण सेट के लेबल को जानते हैं जिसका उपयोग पूर्वानुमान के लिए किया गया था और हमने इस लेबल को अवलोकन लेबल कहा है और पूर्वानुमान के माध्यम से प्राप्त लेबल को पूर्वानुमान लेबल कहा जाता है।

```
observed <- test_set$Species
predicted <- y_pred
#Creating confusion matrix
#install.packages("caret")
library(caret)
conmat <- confusionMatrix(data=predicted, reference = observed)
```

```
> conmat
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         15          0         0
  versicolor      0         12         1
  virginica       0          3        14

Overall Statistics

               Accuracy : 0.9111
                 95% CI : (0.7878, 0.9752)
    No Information Rate : 0.3333
    P-Value [Acc > NIR] : 8.467e-16

                  Kappa : 0.8667

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.8000           0.9333
Specificity                 1.0000            0.9667           0.9000
Pos Pred Value              1.0000            0.9231           0.8235
Neg Pred Value              1.0000            0.9062           0.9643
Prevalence                  0.3333            0.3333           0.3333
Detection Rate              0.3333            0.2667           0.3111
Detection Prevalence        0.3333            0.2889           0.3778
Balanced Accuracy           1.0000            0.8833           0.9167
```

संवेदनशीलता सही ढंग से पूर्वानुमान सकारात्मक मामलों का अनुपात है। विशिष्टता सही ढंग से पूर्वानुमान नकारात्मक उदाहरणों का अनुपात है। सकारात्मक पूर्वानुमानित मान, सही ढंग से पूर्वानुमानित सकारात्मक उदाहरणों की संख्या का पूर्वानुमानित सकारात्मकता की कुल संख्या का अनुपात है, जिसे सटीक के रूप में भी जाना जाता है। इसी तरह, ऋणात्मक पूर्वानुमानित मान सही ढंग से पूर्वानुमानित ऋणात्मक उदाहरणों की संख्या का पूर्वानुमानित नकारात्मकों की कुल संख्या का अनुपात है। डिटेक्शन प्रचलन को पूर्वानुमानित सकारात्मक घटनाओं (सच्ची सकारात्मक और झूठी सकारात्मक दोनों) की संख्या को पूर्वानुमान की कुल संख्या से विभाजित करके परिभाषित किया गया है। पता लगाने की दर कुल उदाहरणों में से सही ढंग से अनुमानित उदाहरणों का अनुपात है। संतुलित सटीकता संवेदनशीलता और विशिष्टता का औसत है। अन्य प्रदर्शन मेट्रिक्स भी हैं जिनकी गणना मशीन लर्निंग एल्गोरिदम के वर्गीकरण प्रदर्शन के मूल्यांकन के लिए की जा सकती है।

आर -में एस.वी.एम. प्रतिगमन लागू करना

एसवीएम वर्गीकरण के समान सिद्धांत के आधार पर, सपोर्ट वेक्टर रिग्रेशन (एसवीआर) श्रेणीबद्ध प्रतिक्रिया के बजाय संख्यात्मक निर्भर चर के लिए उपयोगी है। एसवीआर एक गैर–पैरामीट्रिक तकनीक है जो सरल रैखिक प्रतिगमन के विपरीत, स्वतंत्र और आश्रित चर दोनों के अंतर्निहित वितरण पर निर्भर नहीं करती है। यह अधिकतम मार्जिन के सिद्धांत पर आधारित है जो एसवीआर को उत्तल अनुकूलन समस्या के रूप में देखने की अनुमति देता है। एसवीआर मॉडल की ओवरफिटिंग से बचने के लिए पेनल्टी पैरामीटर (लागत) को भी शामिल किया जा सकता है। आइए चर्चा करें कि नमूना डेटासेट का उपयोग करके आर–पैकेज "ई1071" का उपयोग करके एसवीआर को कैसे फिट किया जाए। यहां, हम लॉस एंजिल्स ओजोन प्रदूषण डेटा, 1976 का उपयोग आर–पैकेज "एमएलबेंच" में करेंगे।

```
library(e1071)
library(mlbench)
library(caret)
library(MLmetrics)
data(Ozone)
```

इस डेटासेट में 13 चरों पर 366 अवलोकन हैं, प्रत्येक अवलोकन एक दिन के आधार पर है

V1	Month (1 = January, ..., 12 = December)
V2	Day of month (1, 2, …,31)
V3	Day of week (1 = Monday, ..., 7 = Sunday)
V4 (y)	Daily maximum one-hour-average ozone reading
V5	500 millibar pressure height (m) measured at Vandenberg AFB
V6	Wind speed (mph) at Los Angeles International Airport (LAX)
V7	Humidity (%) at LAX
V8	Temperature (degrees F) measured at Sandburg, CA
V9	Temperature (degrees F) measured at El Monte, CA
V10	Inversion base height (feet) at LAX
V11	Pressure gradient (mm Hg) from LAX to Daggett, CA
V12	Inversion base temperature (degrees F) at LAX
V13	Visibility (miles) measured at LAX

चर V5, V7, V8, V9, V10, V11 और V12 में "NA" मान होते हैं और इसलिए संबंधित अवलोकन हटा दिए जाते हैं और परिणामी डेटासेट में 13 चर पर 203 अवलोकन शामिल होते हैं।

```
#Removing NA values
dat <- na.omit(Ozone)
rownames(dat)<- as.numeric(1: nrow(dat))
head(dat)
```

```
> head(dat)
  V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111   9 63.14 150
```

तो, यहाँ हमारा उद्देश्य प्रतिक्रिया की पूर्वानुमान करना है, अर्थात दैनिक अधिकतम एक घंटे–औसत ओजोन रीडिंग (y)। सबसे पहले, हम प्रशिक्षण और परीक्षण डेटासेट तैयार करेंगे।

```
set.seed(123)
index <- createDataPartition(dat$V4,   p = .7, list = FALSE)
train <-  dat[index, ]
test  <- dat[-index, ]
```

```
> head(train)
  V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111   9 63.14 150
8  1 14  3  4 5780  6 19 54 56.12 5000 -44 56.30 200
9  1 15  4  4 5830  3 19 58 62.24 1249 -53 75.74 250
> head(test)
   V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
1   1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2   1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
7   1 13  2  5 5760  6 33 51 57.56  492 -44 64.58  40
17  1 30  5 11 5790  3 28 63 57.38  793 -15 65.84 120
20  2  4  3  2 5590  3 76 36 37.40 5000  70 37.94 100
26  2 13  5  6 5700  4 86 55 49.28 2398  21 53.78 200
```

प्रशिक्षण और परीक्षण डेटा सेट तैयार हैं। अब, हम डिफ़ॉल्ट पैरामीटर सेटिंग के साथ प्रशिक्षण डेटासेट का उपयोग करके एसवीएम प्रतिगमन मॉडल को फिट करेंगे। जैसा कि पहले उल्लेख किया गया है, कोई भी प्रशिक्षण और पूर्वानुमान के लिए चार कर्नेल यानी 'रैखिक', 'बहुपद', 'रेडियल आधार' और 'सिग्मॉइड' में से किसी एक कर्नेल फ़ंक्शन को चुन सकता है। यहां, हम "रेडियल" कर्नेल फ़ंक्शन का उपयोग करेंगे जो कि डिफ़ॉल्ट कर्नेल पैरामीटर है।

```
#fitting of the SVM regression model
svr_model = svm(train$V4~., data=train)
summary(svr_model)
```

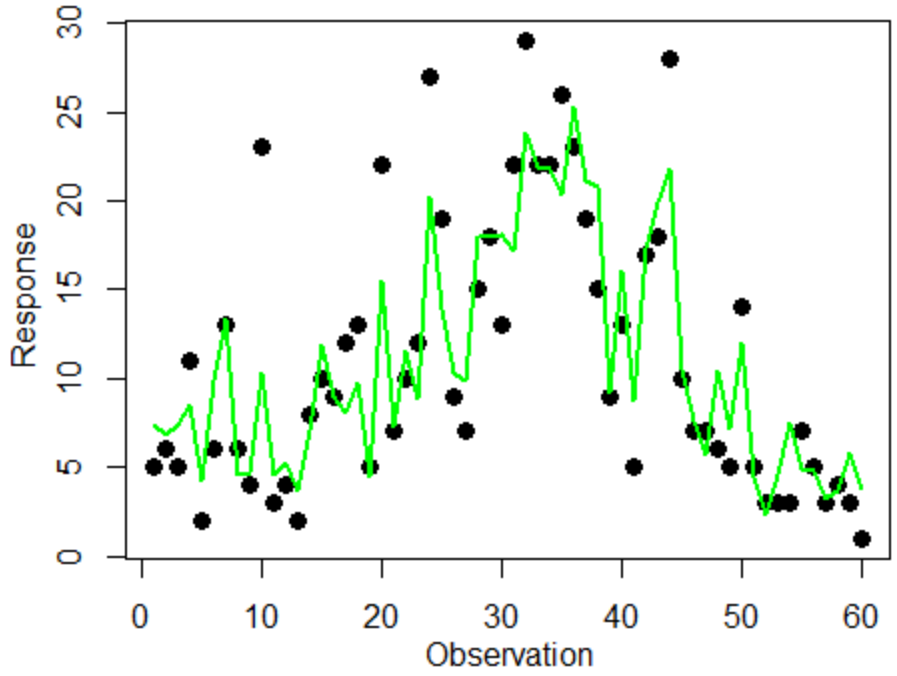
```
Call:
svm(formula = train$V4 ~ ., data = train)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.01754386
    epsilon:  0.1


Number of Support Vectors:  121
```

अगले परीक्षण सेट के लिए पूर्वानुमान करेंगे और वास्तविक मानों के साथ अनुमानित अवलोकन मानों को प्लॉट करेंगे।

```
#Prediction for the test set
pred_svr <- predict(svr_model, test[,-4])
print(pred_svr)
```

```
> print(pred_svr)
        1         2         7        17        20        26
 7.401885  6.873475  7.417459  8.526543  4.230745  9.856104
       28        29        30        34        36        38
13.310310  4.513160  4.687134 10.321620  4.556551  5.226023
       42        43        50        53        54        60
 3.634022  7.074853 11.854427  8.968529  8.094056  9.696712
```

```
#plotting
x <- 1:length(test$V4)
plot(x, test$V4, pch=16, col="black", cex=1.3, xlab="Observation",
ylab="Response")
lines(x, pred_svr, lwd="2", col="green")
```



अब, हम विभिन्न मैट्रिक्स जैसे माध्य वर्ग त्रुटि (एमएसई), माध्य निरपेक्ष त्रुटि (एमएई), मूल माध्य वर्ग त्रुटि (आरएमएसई), आर–वर्ग और माध्य निरपेक्ष प्रतिशत त्रुटि (एमएपीई) के साथ मॉडल के प्रदर्शन (पूर्वानुमान सटीकता) का मूल्यांकन करेंगे।

```
#Performance metrics
mse <-  MSE(y_true=test$V4, y_pred=pred_svr)
mae <- MAE(y_true=test$V4, y_pred=pred_svr)
mape <- MAPE (y_true=test$V4, y_pred=pred_svr)
rmse <- RMSE(y_true=test$V4, y_pred=pred_svr)
Rsqr <- R2_Score(y_true=test$V4, y_pred=pred_svr)
Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse, R2=Rsqr)
Accuracy
```

```
Accuracy
     MSE      MAE      MAPE     RMSE        R2
11.21423 2.471234 0.332201 3.348766 0.8041752
```

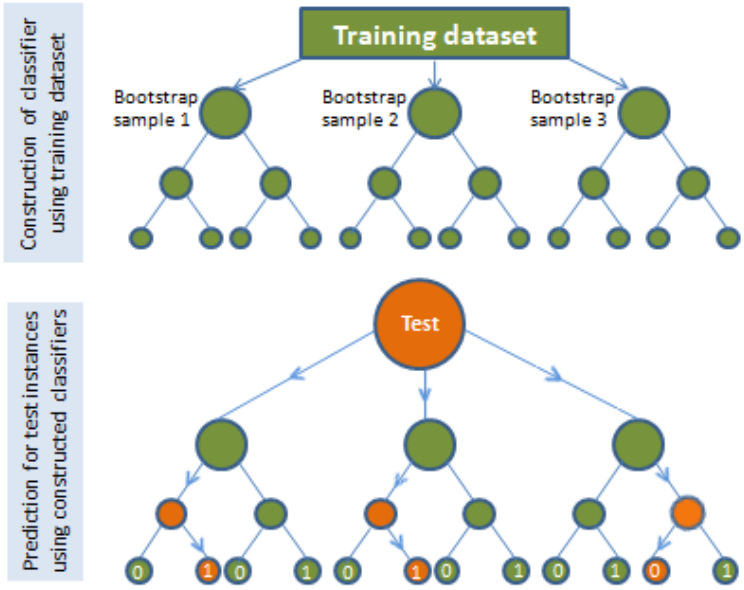
एसवीआर में, कर्नेल फ़ंक्शन के हाइपर पैरामीटर को अनुकूलित करके पूर्वानुमान सटीकता में सुधार किया जा सकता है। यह भी सलाह दी जाती है कि तुलनात्मक सटीकता के बारे में बेहतर विचार रखने के लिए उपयोगकर्ता को विभिन्न कर्नेल फ़ंक्शन की सटीकता की तुलना करनी चाहिए।

रैंडम फ़ॉरेस्ट

वर्गीकरण और प्रतिगमन ट्री (कार्ट) ट्री के प्रत्येक नोड पर सूचना प्राप्त के सिद्धांत पर काम करते हैं। दूसरे शब्दों में, उस नोड पर विभाजन होता है जहां सूचना अधिकतम प्राप्त होती है। यह प्रक्रिया तब तक दोहराई जाती है जब तक कि सभी नोड्स समाप्त नहीं हो जाते या कोई और जानकारी प्राप्त नहीं हो जाती। कार्ट में बहुत कम

पूर्वानुमान करने की शक्ति होती है और इसे अक्सर कमजोर शिक्षार्थी के रूप में संदर्भित किया जाता है। आरएफ एल्गोरिथम ऐसे कमजार शिक्षार्थियों की सामूहिक अवधारणा पर आधारित है।

रैंडम फ़ॉरेस्ट (आरएफ) एक पर्यवेक्षित मशीन लर्निंग एल्गोरिथम है जिसका उपयोग वर्गीकरण और प्रतिगमन समस्याओं दोनों के लिए किया जा सकता है। आरएफ एक समूह में सीखने की विधि है जिसमें कई ट्री–आधारित क्लासिफायरियर शामिल हैं, जहां प्रत्येक क्लासिफायरियर (वर्गीकरण ट्री) का निर्माण प्रशिक्षण डेटासेट के बूटस्ट्रैप पुनः नमूने पर किया जाता है। यह विधि किंवदंती के लिए मजबूत है, अधिक फिटिंग की समस्या है और बड़े डेटासेट को संभाल सकती है। प्रत्येक बूटस्ट्रैप नमूने में, एक वर्गीकरण ट्री का निर्माण किया जाता है और उस ट्री क्लासिफायरियर के लिए परीक्षण सेट के रूप में क्लासिफायर निर्माण में भाग लेने से परहेज करने वाले अवलोकनों का उपयोग किया जाता है। औसतन, आरएफ में प्रत्येक क्लासिफायरियर प्रशिक्षण डेटा के 2/3 पर बनाया गया है और 1/3 आउट–ऑफ–बैग नमूने पर परीक्षण किया गया है। ये ओओबी नमूने आरएफ की पूर्वानुमान त्रुटि को मापने के लिए डेटा के स्रोत हैं। अधिक स्पष्ट रूप से, आरएफ में प्रत्येक क्लासिफायरियर के लिए त्रुटि को उसके ओओबी नमूनों (ओओबी त्रुटि के रूप में कहा जाता है) के आधार पर मापा जाता है और इन ओओबी त्रुटियों को फ़ॉरेस्ट के ओओबी त्रुटि की गणना करने के लिए सभी निर्णय ट्री पर औसत किया जाता है। जहां तक परीक्षण उदाहरण के वर्गीकरण का संबंध है, आरएफ के प्रत्येक वर्गीकरणकर्ता प्रत्येक परीक्षण उदाहरणों को पूर्व–निर्धारित वर्गों में से एक के लिए वोट देते हैं और परीक्षण उदाहरण की पूर्वानुमान विजेता वर्ग के लेबल द्वारा की जाती है। आरएफ प्रतिगमन के संबंध में, जंगल में प्रत्येक व्यक्तिगत निर्णय ट्री द्वारा की गई पूर्वानुमान का औसत लेकर अंतिम पूर्वानुमान की जाती है। आरएफ भी प्रकृति में गैर–पैरामीट्रिक है जो आश्रित और स्वतंत्र चर के सांख्यिकीय वितरण के बारे में अंतर्निहित धारणा पर निर्भर नहीं करता है।



छवि स्रोतः मेहर एट अल। (2019) बीएमसी जेनेटिक्स 20 (1), 1–13

आर में रैंडम फ़ॉरेस्ट वर्गीकरण लागू करना

आरएफ वर्गीकरण को लागू करने के लिए, पहले हमें रैंडम फ़ॉरेस्ट आर –पैकेज स्थापित करना होगा।

```
install.packages(randomForest)
library(randomForest)
```

यहां, हम आर यानी आईरिस डेटासेट में उपलब्ध इनबिल्ट डेटासेट का उपयोग करेंगे। यह एक बेंचमार्क डेटासेट है जिसमें तीन अलग–अलग पौधों की प्रजातियों के 150 अवलोकन शामिल हैं, प्रत्येक प्रकार के लिए 50 अवलोकनों के साथ ई सेटोसा, वर्सिकलर और वर्जिनिका चार चर (पूर्वानुमानक) हैं जैसे सीपल लंबाई, सेपल चौड़ाई, पंखुड़ी लंबाई और पंखुड़ी चौड़ाई। यह डेटा केवल आईरिस टाइप करके आर कंसोल में लोड किया जा सकता है।

```
#Load the dataset
dat <- iris
```

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width    Species
1            5.1         3.5          1.4         0.2     setosa
2            4.9         3.0          1.4         0.2     setosa
3            4.7         3.2          1.3         0.2     setosa
4            4.6         3.1          1.5         0.2     setosa
5            5.0         3.6          1.4         0.2     setosa
```

वर्गीकरण या प्रतिगमन के लिए, प्रशिक्षण और परीक्षण डेटासेट की आवश्यकता होती है। प्रशिक्षण सेट में संबंधित लेबल के साथ अवलोकन होते हैं जबकि परीक्षण सेट के अवलोकन के लिए लेबल की पूर्वानुमान की जाती है। हम डेटासेट को ट्रेन और सत्यापन सेट में 70:30 के अनुपात में विभाजित करेंगे। प्रशिक्षण और सत्यापन के लिए उपयोग किए जाने वाले डेटासेट का प्रतिशत उपयोगकर्ता पर निर्भर करता है।

```
# Split into Train and Validation sets
# Training Set : Validation Set = 70 : 30 (random)
set.seed(100)
train <- sample(nrow(dat), 0.7*nrow(dat), replace = FALSE)
Train_Set <- dat[train,]
Valid_Set <- dat[-train,]
summary(Train_Set)
summary(Valid_Set)
```

```
> summary(Train_Set)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.200   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.300   Median :1.300
 Mean   :5.811   Mean   :3.054   Mean   :3.702   Mean   :1.197
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
       Species
 setosa    :36
 versicolor:34
 virginica :35
```

```
> summary(Valid_Set)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.400   Min.   :2.000   Min.   :1.300   Min.   :0.100
 1st Qu.:5.300   1st Qu.:2.800   1st Qu.:1.500   1st Qu.:0.200
 Median :5.900   Median :3.000   Median :4.400   Median :1.400
 Mean   :5.918   Mean   :3.064   Mean   :3.889   Mean   :1.204
 3rd Qu.:6.500   3rd Qu.:3.300   3rd Qu.:5.500   3rd Qu.:1.800
 Max.   :7.700   Max.   :4.200   Max.   :6.700   Max.   :2.500
       Species
 setosa    :14
 versicolor:16
 virginica :15
```

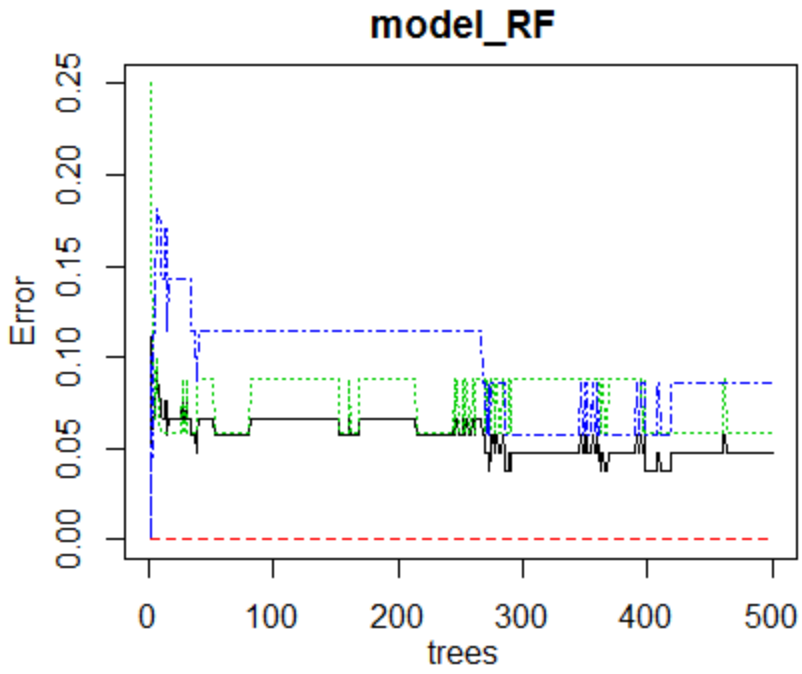
अब, हम डिफ़ॉल्ट पैरामीटर सेटिंग के साथ आरएफ मॉडल को प्रशिक्षित करेंगे। आरएफ मॉडल में ट्यून किए जाने के लिए मुख्य रूप से दो पैरामीटर हैं, यानी बढ़ने के लिए ट्री की संख्या (ntree) और प्रत्येक विभाजन (mtrey) पर बेतरतीब ढंग से नमूने लिए गए चर की संख्या। ntree मान को दो छोटे सेट नहीं किया जाना चाहिए। यह सुनिश्चित किया जाना चाहिए कि प्रत्येक अवलोकन की पूर्वानुमान कम से कम कुछ बार हो। के लिए डिफ़ॉल्ट mtrey मान चर की संख्या का वर्गमूल है और प्रतिगमन समस्या के लिए पूर्वानुमान की संख्या का एक तिहाई है। n tree का डिफ़ॉल्ट मान 500 है।

```
# Create a Random Forest model with default parameters
model_RF <- randomForest(Species ~ ., data = Train_Set)
model_RF
```

```
> model_RF

Call:
 randomForest(formula = Species ~ ., data = Train_Set)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 2

        OOB estimate of  error rate: 4.76%
Confusion matrix:
           setosa versicolor virginica class.error
setosa         36          0         0  0.00000000
versicolor      0         32         2  0.05882353
virginica       0          3        32  0.08571429
```

यहां, हम देख सकते हैं कि वर्गीकरण ट्री की संख्या 500 है और प्रत्येक विभाजन पर आजमाए गए चरों की संख्या 2 है जो 4 का वर्गमूल है। समग्र ओओबी त्रुटि 4.76% है।

सेटोज़, वर्सिकलर और वर्जिनिका के लिए गलत वर्गीकरण त्रुटियाँ क्रमशः 0, 0.05 और 0.08 हैं। कोई भी आरएफ मॉडल को प्लाट सकता है और ट्री की संख्या के संबंध में त्रुटि दर की कल्पना कर सकता है।



अब हम प्रशिक्षण सेट के साथ–साथ सत्यापन सेट के लिए भी पूर्वानुमान करेंगे।

```
# Predicting on train set
pred_Train <- predict(model_RF, Train_Set[,-5], type = "class")
# Checking classification accuracy
table(pred_Train, Train_Set$Species)
```

```
> table(pred_Train, Train_Set$Species)

pred_Train    setosa versicolor virginica
  setosa          36          0         0
  versicolor       0         34         0
  virginica        0          0        35
```

```
# Predicting on Validation set
pred_Valid <- predict(model_RF, Valid_Set[,-5], type = "class")
# Checking classification accuracy
table(pred_Valid,Valid_Set$Species)
```

```
> table(pred_Valid,Valid_Set$Species)

pred_Valid    setosa versicolor virginica
  setosa          14          0         0
  versicolor       0         14         0
  virginica        0          2        15
```

प्रशिक्षण सेट की पूर्वानुमान करते समय, हम देख सकते हैं कि तीनों प्रजातियों के सभी उदाहरणों को सही ढंग से वर्गीकृत किया गया है। यह इस तथ्य के कारण हो सकता है कि इस मामले में प्रशिक्षण और परीक्षण सेट दोनों समान हैं। दूसरी ओर, सेटोसा और वर्सिकलर के सभी परीक्षण उदाहरणों की सही पूर्वानुमान की जाती है, जबकि वर्जिनिका के 2 अवलोकनों को वर्सिकलर में गलत वर्गीकृत किया जाता है। इसलिए, परीक्षण की सटीकता हमेशा प्रशिक्षण सेट के बराबर या उससे कम होगी। अगला कदम भ्रम मैट्रिक्स और उसके बाद प्रदर्शन मीट्रिक की गणना

करना है। सबसे पहले, हम परीक्षण सेट के लेबल को जानते हैं जिसका उपयोग पूर्वानुमान के लिए किया गया था और हमने इस लेबल को मनाया लेबल कहा है और पूर्वानुमान के माध्यम से प्राप्त लेबल को पूर्वानुमान लेबल कहा जाता है।

```
observed <- Valid_Set$Species
predicted <- pred_Valid
#Creating confusion matrix
#install.packages("caret")
library(caret)
conmat <- confusionMatrix(data=predicted, reference = observed)
conmat
```

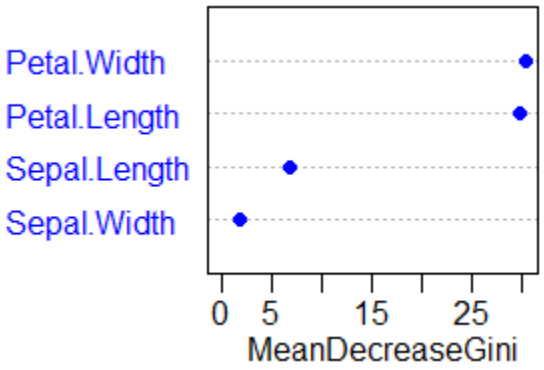
```
> conmat
Confusion Matrix and Statistics

            Reference
Prediction   setosa versicolor virginica
  setosa         14          0         0
  versicolor      0         14         0
  virginica       0          2        15

Overall Statistics

               Accuracy : 0.9556
                 95% CI : (0.8485, 0.9946)
    No Information Rate : 0.3556
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9333

 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: setosa Class: versicolor Class: virginica
Sensitivity                 1.0000            0.8750           1.0000
Specificity                 1.0000            1.0000           0.9333
Pos Pred Value              1.0000            1.0000           0.8824
Neg Pred Value              1.0000            0.9355           1.0000
Prevalence                  0.3111            0.3556           0.3333
Detection Rate              0.3111            0.3111           0.3333
Detection Prevalence        0.3111            0.3111           0.3778
Balanced Accuracy           1.0000            0.9375           0.9667
```

हम प्रत्येक चर के चेक महत्व का भी उपयोग कर सकते हैं। प्रत्येक चर के लिए सटीकता में औसत कमी की गणना निम्नलिखित फ़ंक्शन का उपयोग करके की जा सकती है और प्लॉट की जा सकती है।

variables can be computed and plotted by using the following function.

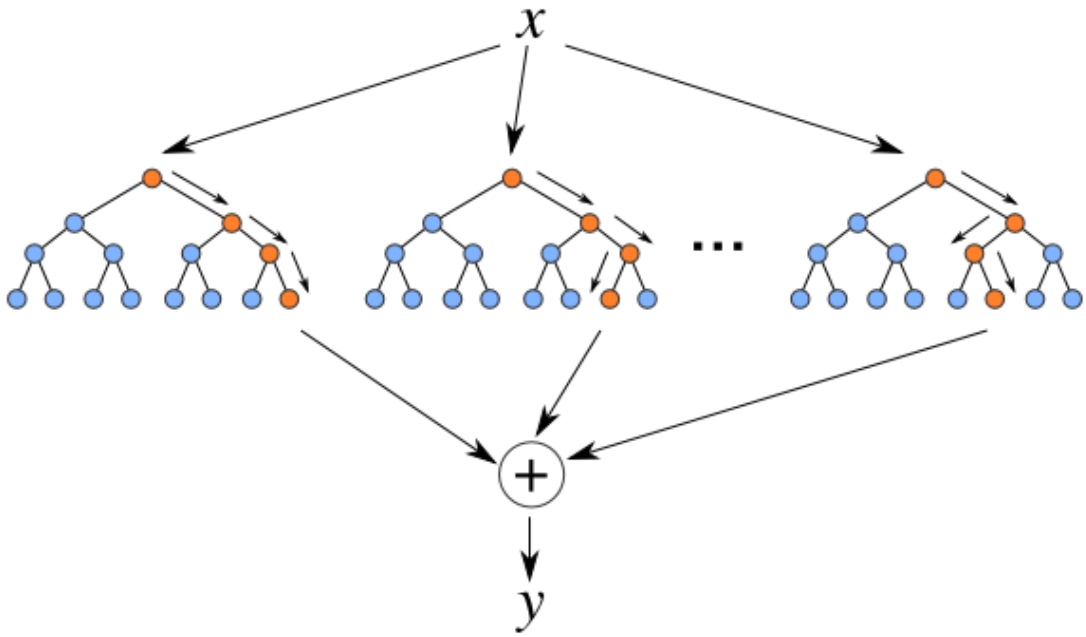
```
importance(model_RF)#computation
varImpPlot(model_RF)#plotting
```

```
> importance(model_RF)
             MeanDecreaseGini
Sepal.Length         6.923897
Sepal.Width          1.894913
Petal.Length        29.887463
Petal.Width         30.505017
```



आर में रैंडम फ़ॉरेस्ट रिग्रेशन लागू करना

सरल रैखिक प्रतिगमन की तरह, आरएफ प्रतिगमन आश्रित और स्वतंत्र चर की अवधारणा पर आधारित है। आरएफ रिग्रेशन में, एन्सेम्बल लर्निंग तकनीक कार्यरत है जो कई ट्री–आधारित शिक्षार्थियों के परिणामों को जोड़ती है। कई अन्य प्रतिगमन विधियों की तुलना में आरएफ प्रतिगमन अधिक सटीक और शक्तिशाली है। यह उन डेटासेट पर भी अच्छा प्रदर्शन करता है जिनमें गैर–रैखिक संबंध वाली विशेषताएं हैं। हालांकि, ओवरफिटिंग से बचने के लिए पर्याप्त संख्या में ट्री का निर्माण किया जाना चाहिए।



छवि स्रोतः https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

अब, हम आर –सॉफ़्टवेयर का उपयोग करके आरएफ समाश्रयण करने के लिए विभिन्न चरणों पर चर्चा करेंगे। तो, पहले हमें आर –कन्सोल में में रैंडम फ़ॉरेस्ट आर–पैकेज को स्थापित करने की आवश्यकता है।

```
#package installation
install.packages("randomForest")
library(randomForest)
```

यहां, हम उसी डेटासेट का उपयोग करेंगे जिसका उपयोग आर में सपोर्ट वेक्टर रिग्रेशन को लागू करने के लिए किया गया है। दूसरे शब्दों में, हम लॉस एंजिल्स ओजोन प्रदूषण डेटा, 1976 का उपयोग आर–पैकेज "एमएलबेंच" में करेंगे। डेटासेट के बारे में अधिक विवरण "आर में समर्थन वेक्टर प्रतिगमन को लागू करना" उप–अनुभाग में पाया जा सकता है। यहां भी, हम एनए मानों को हटाने के बाद डेटासेट का उपयोग करेंगे।

```
#Loading the dataset
library(mlbench)
data(Ozone)

#Removing NA values
dat <- na.omit(Ozone)
rownames(dat)<- as.numeric(1: nrow(dat))
head(dat)
```

```
> head(dat)
  V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
1  1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2  1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111   9 63.14 150
```

तो, यहाँ हमारा उद्देश्य प्रतिक्रिया की पूर्वानुमान करना है, अर्थात दैनिक अधिकतम एक घंटे–औसत ओजोन रीडिंग (y)। सबसे पहले, हम पहले प्रशिक्षण और परीक्षण डेटासेट तैयार करेंगे। प्रशिक्षण और परीक्षण सेट में डेटासेट का विभाजन इस मायने में महत्वपूर्ण है कि प्रशिक्षण सेट में प्रतिक्रिया और पूर्वानुमानक दोनों होते हैं जिनसे मॉडल सीखता है। परीक्षण सेट तब प्रशिक्षण सेट से सीखे गए मॉडल के आधार पर मॉडल की पूर्वानुमान का परीक्षण करता है।

```
set.seed(123)
index <- createDataPartition(dat$V4,   p = .7, list = FALSE)
train <-  dat[index, ]
test  <- dat[-index, ]
```

```
> head(train)
  V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
3  1  7  3  4 5790  6 19 45 46.40 2631 -33 54.14 100
4  1  8  4  4 5790  3 25 55 52.70  554 -28 64.76 250
5  1  9  5  6 5700  3 73 41 48.02 2083  23 52.52 120
6  1 12  1  6 5720  3 44 51 54.32  111   9 63.14 150
8  1 14  3  4 5780  6 19 54 56.12 5000 -44 56.30 200
9  1 15  4  4 5830  3 19 58 62.24 1249 -53 75.74 250
> head(test)
   V1 V2 V3 V4   V5 V6 V7 V8    V9  V10 V11   V12 V13
1   1  5  1  5 5760  3 51 54 45.32 1450  25 57.02  60
2   1  6  2  6 5720  4 69 35 49.64 1568  15 53.78  60
7   1 13  2  5 5760  6 33 51 57.56  492 -44 64.58  40
17  1 30  5 11 5790  3 28 63 57.38  793 -15 65.84 120
20  2  4  3  2 5590  3 76 36 37.40 5000  70 37.94 100
26  2 13  5  6 5700  4 86 55 49.28 2398  21 53.78 200
```

प्रशिक्षण और परीक्षण डेटा सेट तैयार हैं। अब, हम mtry और ntree मापदंडों के डिफ़ॉल्ट मानों के साथ प्रशिक्षण डेटासेट का उपयोग करके आरएफ प्रतिगमन मॉडल को फिट करेंगे। जैसा कि पहले उल्लेख किया गया है, अधिकतम सटीकता प्राप्त करने के लिए कोई भी mtry और ntree को अनुकूलित कर सकता है। यहाँ, हमने mtry=4 (पूर्वानुमान की संख्या का एक तिहाई) और ntree=500 (डिफ़ॉल्ट मान) का उपयोग किया।

```
#fitting of the RF regression model
RF_model = randomForest(train$V4~., data=train)
summary(RF_model)
```

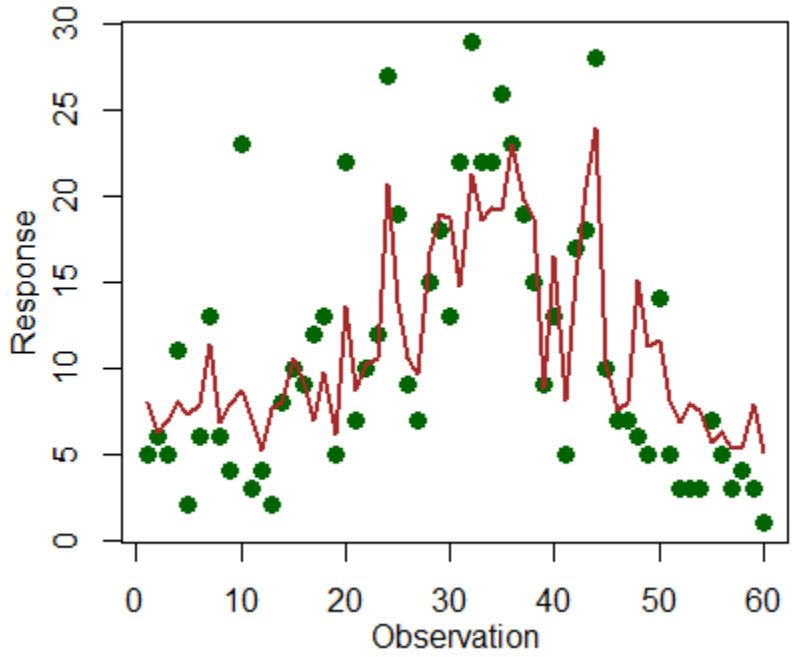
```
> RF_model

Call:
 randomForest(formula = train$V4 ~ ., data = train)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 4

          Mean of squared residuals: 24.82561
                    % Var explained: 64.86
```

अगला, हम परीक्षण सेट के लिए पूर्वानुमान करेंगे और वास्तविक मूल्यों के साथ अनुमानित अवलोकन मूल्यों को प्लॉट करेंगे।

```
#Prediction for the test set
pred_RF <- predict(RF_model, test[,-4])
print(pred_RF)
```

```
> print(pred_RF)
        1         2         7        17        20        26
 7.960000  6.265333  6.999900  7.999733  7.343100  7.855833
       30        34        36        38        42        43
 7.900767  8.737433  7.077767  5.209833  7.661100  7.970033
       54        60        61        68        70        71
 6.946400  9.707567  6.120767 13.533100  8.640700 10.195667
       88        90        91       100       103       107
13.788700 10.557533  9.675800 16.693967 18.934533 18.772033
```

```
#plotting
x <- 1:length(test$V4)
plot(x, test$V4, pch=16, col="darkgreen", cex=1.3, xlab="Observation",
ylab="Response")
lines(x, pred_RF, lwd="2", col="brown")
```



अब, हम विभिन्न मैट्रिक्स जैसे माध्य वर्ग त्रुटि (एमएसई), माध्य निरपेक्ष त्रुटि (एमएई), मूल माध्य वर्ग त्रुटि (आरएमएसई), आर–वर्ग और माध्य निरपेक्ष प्रतिशत त्रुटि (एमएपीई) के साथ मॉडल के प्रदर्शन (पूर्वानुमान सटीकता) का मूल्यांकन करेंगे।

```
#Performance metrics
library(MLmetrics)
mse <-  MSE(y_true=test$V4, y_pred=pred_svr)
mae <- MAE(y_true=test$V4, y_pred=pred_svr)
mape <- MAPE (y_true=test$V4, y_pred=pred_svr)
rmse <- RMSE(y_true=test$V4, y_pred=pred_svr)
Rsqr <- R2_Score(y_true=test$V4, y_pred=pred_svr)
Accuracy <- data.frame(MSE=mse, MAE=mae, MAPE=mape, RMSE=rmse, R2=Rsqr)
Accuracy
```

```
Accuracy
     MSE      MAE       MAPE     RMSE        R2
17.35143 3.188692 0.5369437 4.165505 0.6970064
```

आरएफ प्रतिगमन में, पूर्वानुमान सटीकता उपयोग किए गए डेटासेट के प्रकार पर निर्भर करती है। साथ ही, मापदंडों को अनुकूलित करके सटीकता में सुधार किया जा सकता है। यह भी सलाह दी जाती है कि तुलनात्मक सटीकता के बारे में बेहतर विचार रखने के लिए उपयोगकर्ता को विभिन्न प्रतिगमन विधियों की सटीकता की तुलना करनी चाहिए।

संदर्भ

ब्रिमन, एल।, 2001. रैंडम फ़ॉरेस्ट । मशीन लर्निंग, 45(1), 5–32.

लियाव, ए। एन्ड वीनर, एम।,2002। क्लासिफिकेशन एन्ड रिग्रेशन बाय रैंडमफॉरेस्ट। आर न्यूज, 2(3),18–22.

मेयर, डी., दिमित्रीडौ, ई., हॉर्निक, के., वेइंगसेल, ए., लीश, एफ., चांग, सी.सी., लिन, सी.सी. और मेयर, एम.डी., 2019। पैकेज 'ई1071'। आर जर्नल।

वापनिक, वी।,गयोन, आई। और हेस्टी, टी।, 1995। सपोर्ट वेक्टर रिग्रेशन। मशीन लर्निंग, 20(3), 273–297।

आनुवंशिकता के आकलन करने के लिए विभिन्न तरीकों की तुलना

ए के पॉल

भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012

amrit.paul@icar.gov.in

पशु प्रजनन में ज्यादातर द्विगुण विशेषक (Binary Traits) उत्पादन दक्षता(efficiency) के महत्वपूर्ण निर्धारक(Determinants)होते हैं और मूलभूत कारकों के सूचक हैं जिनकी माप मुश्किल या मंहगी होती हैं। लक्षणों मामले में जिनका दृश्यप्रारूप(phenotype)सामान्य संस्थापक(classical) विधियों द्वारा अभिव्यक्त किया जाता है। यह सीधे प्रयुक्त नहीं किये जाते।वह लक्षण जिनकी वंशागति बहु–उपादानीय (multifactorial) होती है लेकिन जिनमें किसी भी प्रकार नहीं या सभी प्रकार की दृश्य प्रारूप अभिव्यक्त होते हैं। वह प्रारम्भिक लक्षण कहलाते हैं।एक ऐसा ही लक्षण जो पशु डेरी में प्रारम्भिक की तरह वर्गीकृत है वह गाय के झुण्ड में स्टेएबिलिटीहै। अगर गाय रखी जाती है तो यह माना जाता है कि वह पशुओ केझुण्ड में रहेगी अन्यथा अलग कर दी जायेगी।या तो किसी प्रकार नहीं या सभी प्रकार की विशेषक स्टेएबिलिटी की वंशागतित्व के अनुमान को सरल बनाने के लिए प्रारम्भिक मॉडल मान लिया जाता है। प्रारम्भिक लक्षणों की वंशागतित्व के अनुमानकी अनेक विधियाँ है लेकिन सभी विधियाँ असंतुलित आँकड़े समुच्चय के लिए सीधे तरीके से प्रयुक्त नहीं होती है।इस दृष्टिकोण के लिए डैम्पस्टर लरनर की दोनों बीटा–द्विपद विधियों की असंतुलित आँकडों के मामले में तुलनात्मक निष्पादन के अध्ययन का प्रयास किया गया है।

आँकडें मॉडल

इससे पहले की दो प्रक्रियाओं की संकल्पनाओं पर विचार विमर्श किया जाये, स्टेएबिलिटीके आँकड़े इस प्रकार है। दी हुयी जनसंख्या मेंप्रक्रिया को मानकीकृत गोसियन विचर(Z) के द्वारा समझाया गया है। जिसका माध्य शून्य है तथा प्रसरण एक है। जब भी z एक प्रारम्भिक संख्या से ज्यादा हो जाता है। तब उसे z' मानेंगे, जोकि ज्ञात है। तब एक बाध्य आवलोकन लक्षण ()अभिव्यक्त किया जाता है। यह लक्षण का मान उपस्थिति पर एक तथा अनुपस्थिति पर शून्य है।

अनुपालनीय विचर (z) के लिए रखीय मॉडल

$$Z_{ijk} = \mu + S_i + e_{ijk} \qquad \text{.....(1)}$$

जहाँ z_{ijk} j^{th}ब्लॉक के i^{th}परिवार के k^{th}व्यक्तिगत पर अवलोकन है।

μसम्पूर्ण माध्य है।

s_i i^{th} परिवार प्रभाव है।

e_{ijk}अवशिष्ट प्रभाव है जोकि पिलोटब्लॉक और त्रुटि प्रभाव से गठित है।

$$S_i \sim N(0, \sigma_f^2) \text{vkSj} e_{ijk} \sim N(0, \sigma_e^2)$$

वास्तविक विचर(intrinsic variable)का बाह्य स्कैल(outward scale) पर द्विपद विशेषक (δ)में रूपान्तरण इस प्रकार किया गया।

$$\delta_{ijk} = 1 \quad for \quad Z_{ijk} \leq Z' \; or \; \Phi(Z_{ijk}) \leq P$$

$$= 0 \quad for \quad Z_{ijk} > Z' \; or \; \Phi(Z_{ijk}) > P$$

जहाँ ϕएक सामान्य वितरण संचीय प्रायिकता फलन को दर्शाता है तथा p डाइकोटोमस लक्षण(δ) की निरीक्षण की जनसंख्या प्रायिकता बताता है।

डैमपस्टर और लरनर विधि

डैमपस्टर और लरनर[2] ने द्विगुण विशेषक के लिए व्यक्तिगत संकीर्ण संवेदी वंशागतित्व का अनुमान दिया जोकिh^2_{DL} द्वारा निर्दिष्ट किया गया है। जिसको ज्ञायानोला [4]ने अधिक सामान्य समाधान(Solution)के विशेष मामले की तरह दिखाया है।
जैसे

$$\hat{h}^2_{DL} = 4\hat{\sigma}^2_f(\delta) \times \left[\phi(Z')\right]^{-2}$$

...(2)

जहाँϕ द्विगुण स्कैल $[Z' = \varphi^{-1}(p)]$ पर व्यंजक के लिए प्रारम्भिक पर मूल्यांकित किया गया जो गोसियन प्रायिकता घनत्व फलन को दर्शाता है तथा $\sigma_i^2(f)$ परिवार प्रसरण घटक का अनुमानहै जोकि द्विगुण विशेषक पर प्रयोग किया गया प्रसरण विधि (ANOVA)के विश्लेषण से प्राप्त किया गया है।

<u>बीटा–द्विपद मॉडल प्रस्ताव</u>

निम्नलिखित मैगनूसेन और बरेमर[5] बीटा–प्राचल के तीन समुच्चयः एक दृश्य प्रारूप (phenotypic) परिवार प्रायिकताओं के लिए, एक परिवार प्रायिकताओं के लिए तथा एक संयोजी (additive) जननिक प्रायिकता के लिए δ_{ijk} एक द्विगुण–विशेषक आँकड़ो के मॉडल पर आधारित, वंशागतित्व अनुमानो पर आधारित प्राप्त करने के लिए कल्पित है।

$$p_{\mathrm{ijk}} = p + p_i + p_{\mathrm{ijk}} \qquad \text{...(3)}$$

जहाँ piपरिवार के j^{th}ब्लॉक में k^{th}व्यक्तिगत पर द्विगुण विशेषक *(δ)* की निरीक्षणता की प्रायिकता है, pसम्पूर्ण जनसंख्या प्रायिकता(स्थिर प्रभाव) है और शेष pपरिवार प्रभाव तथा अवशिष्ट के क्रमशः रेन्डम योगदान है। इस मॉडल दृश्य प्रारूप [(pf), परिवार (f) और संयोजी जननिक(a)]से तीनों प्रसरण घटक σ_{pf}^2*(d)*, σ_f^2 *(δ)*और σ_a^2 *(δ)* द्विगुण आँकड़ों *(δ_{ijk})* पर किये गये प्रसरण एक तरफा विश्लेषण द्वारा प्राप्त किया जाता है।

दृश्य प्रारूप परिवार प्रायिकतायें$[p_{(pf)i}]$ एक बीटा–वितरण का अनुसरण करता है।

$$p_{(pf)i} = \sum_{jk} \frac{P_{ijk}}{n_{i\cdot\cdot}} \quad \mathrm{lkFk} \quad \mathrm{p_{(pf)i}} \sim \textit{Beta}\ (\alpha_{pf}, \beta_{pf}) \qquad \text{...(4)}$$

जहाँ $n_{i..}$परिवार iमें की गयी अवलोकनों की संख्या है उसी तरह परिवार प्रायिकतायें सम्पूर्ण माध्य तथा संयोजित (aaditive)परिवार प्रभाव एक योग(sum)द्वारा परिभाषित (defined)की जाती है

$$p_{(f)i} = p + p_i \quad \text{with } p_{(f)i} \sim \textit{Beta}\ (\alpha_f, \beta_f) \qquad \text{.....(5)}$$

यह मानते हुये कि परिवार में हॉफ–सिब है, निम्नलिखित संकलपना–संबंधी (conceptual) मॉडल संयोजी जननिक परिवार प्रायिकताओं$\left[p_{(a)i}\right]$ के लिए प्रयोग होता है।

$$p_{(a)i} = p + 0.5p_i \text{ with } p_{(a)i} \sim \textit{Beta}\ (\alpha_a, \beta_a) \qquad \alpha_a, \beta_a > 0 \qquad \text{....(6)}$$

उपरोक्त प्रायिकताओं के नमूना अनुमानो आँकड़ों से इस प्रकार प्राप्त किये गये हैं।

$$\hat{p}_{(pf)i} = \sum_{jK} \frac{\delta_{ijk}}{n_{i\cdot\cdot}} = \hat{p}_{(f)i} \qquad \text{.....(7)}$$

$$\bar{p} = \sum_i \frac{\hat{p}_{(pf)i}}{n_{fam}} \qquad \text{.....(8)}$$

निम्नलिखित जोनशन एण्ड कौनज[5], तीन समुच्चयों के प्राचलो का अनुमान निम्नलिखित तरीकों से प्राप्त किया जा सकता है बीटा–वितरण का परिवार निम्नलिखित तरह के प्रायिकता घनत्व फलन के सभी वितरणों से प्रकृतिस्थ (composed) है।

$$P_{y(Y)} = \frac{1}{B(\alpha, \beta)} \frac{(Y - a)^{\alpha-1}(b - y)^{\beta-1}}{(b - a)^{\alpha+\beta-1}} \qquad (a \leq y \leq b),\ \alpha, \beta > 0 \qquad \text{..... (9)}$$

वितरण (4.7)में सभी चारों प्राचलो का अनुमान नमूनों की बराबरी करके तथा पहले चार आघूर्ण (moment) की जनसंख्या के मान से प्राप्त किया जा सकता है।

यदि a और b के मान ज्ञात है तब पहला और दूसरा आघूर्ण इस प्रकार दिये जाते हैं।

$$\mu_1' = \frac{a + (b - a)\alpha}{\alpha + \beta} \qquad \text{.....(10)}$$

$$\mu_2 = (b - a)^2 \alpha\beta(\alpha + \beta)^{-2}(\alpha + \beta + 1)^{-1} \qquad \text{....(11)}$$

जहाँ

$$\frac{\mu_1'-a}{b-a} = \frac{\alpha}{\alpha+\beta} \quad \text{and} \quad \frac{\mu_2}{(b-a)^2} = \frac{\alpha}{\alpha+\beta}\left(1-\frac{\alpha}{\alpha+\beta}\right)\frac{1}{\alpha+\beta+1} \qquad \text{.....(12)}$$

तो अब (Thus)

$$\alpha + \beta = \frac{u_1'-a}{b - a} \frac{\left(1 - \frac{u_1'-a}{b - a}\right)}{\left(\frac{\mu_2}{(b - a)^2}\right)} - 1 \qquad \text{...(13)}$$

$$\alpha = \left(\frac{u_1'-a}{b - a}\right)^2 \left(1 - \frac{u_1'-a}{b - a}\right) \left(\frac{\mu_2}{(b - a)^2}\right)^{-1} - \frac{u_1'-a}{b - a} \qquad \text{...(14)}$$

a को 'kwU;और b को एक लेकर उपरोक्त समीकरण घटकर कुछ इस प्रकार है

$$\alpha = \mu_1^2, \frac{(1 - \mu_1')}{\mu_2} - \mu_1' \qquad \text{...(15)}$$

$$\alpha + \beta = \mu_1' \frac{(1 - \mu_1')}{\mu_2} - 1 \qquad \text{...(16)}$$

अब समीकरण (15) और (16) को हल करके तथा मान $\mu_1' = \bar{P}$ और$\mu_2 = \sigma_t^2$को रखते हुये तीन समुच्चोकाअनुमान सामान्य तरीके में इस प्रकार लिखा जा सकता है

$$\hat{a}_t = \bar{P}^2 \times \frac{(1-\bar{P})}{\hat{\sigma}_t^2(\delta)} - \bar{P}$$ ...(17)

$$\hat{\beta}_t = \frac{(1-\bar{P})}{\bar{P}} \hat{\alpha}_t$$ ...(18)

जहाँ पर सब्सक्रिप्टtप्राचल[t=f (परिवार), pf (दृश्य प्रारूप परिवार माध्य)] या परिवार संयोजी जननिक)] के प्रकार को दर्शाता है और $\hat{\sigma}_t^2(\delta)$ अनुकरणीत द्विगुण–विशेषक (δ) के प्रसरण के विश्लेषण से अनुमानित समरूप प्रसरण घटक है।

सरले इत्यादिका अनुसरण करते हुये प्रतिबंधित माध्य परिवार प्रायिकता चुनी हुयी जनसंख्या में कुछ इस प्रकार अपेक्षा की जाती है।

$$\bar{p}_{t/\delta} = \frac{\delta + \hat{\alpha}_t}{1 + \hat{\alpha}_t + \hat{\beta}_t}$$ ... (19)

जहाँ t=(f,pf,a) सोच–विचार के अन्तर्गत प्रभाव को दर्शाता है माध्य में चयन के कारण बदलाव $(p_{t/\delta} - \bar{p})$ को चयन पर अनुक्रिया (प्रतिक्रिया) माना जा सकता है। जिससे चयन के अन्तर्गत विशेषक की अपेक्षा उपलब्धी वंशागतित्व का अनुमान किया जा सकता है।

चयन प्रतिक्रिया का अनुमान

$$\Phi^{-1}(p_{t/\delta}) - \Phi^{-1}(\bar{p})$$ ...(20)

उपलब्ध (Realized)व्यक्तिगत संकीर्ण संवेदी वंशागतित्व का बीटा–द्विपदअनुमान इस प्रकार है

$h^2 =$<u>चयन पर आपेक्षित प्रतिक्रिया</u>
दृश्य प्रारूपित चयन विभेदन ...

बीटा वितरण प्राचल संयोजी परिवार प्रसरण तथा दृश्य प्रारूप परिवार प्रसरण का अनुपात लेते हुये परिवार माध्यवंशागतित्व को संगणक करने के लिये प्रयोग किया जा सकता है।

$$h^2_{f\,(beta)} = \frac{\hat{\alpha}_f \times \hat{\beta}_f \times \left(\hat{\alpha}_{pf} + \hat{\beta}_{pf}\right)^2 \times \left(\hat{\alpha}_{pf} + \hat{\beta}_{pf} + 1\right)}{\hat{\alpha}_{pf} \times \hat{\beta}_{pf} \times \left(\hat{\alpha}_f + \hat{\beta}_f\right)^2 \times \left(\hat{\alpha}_f + \hat{\beta}_f + 1\right)}$$ ...(22)

प्राचल a और b के साथ एक बीटा वितरण का प्रसरण है।

$$\frac{ab}{(a+b+1)(a+b)^2}$$

बीटा–द्विपद मॉडल में जड़ित परिवार माध्य वंशागतित्व का एक वैकल्पिक फार्मूला परिवार माध्य प्रायिकता$(\bar{p}_{f/\delta} - p)$ में उपलब्ध चयन प्रतिक्रिया तथा दृश्य प्रारूप परिवार माध्य लैवल $(\bar{p}_{f/\delta} - p)$ चयन प्रतिक्रिया के अनुपात

से प्राप्त किया जाता है। संचीय वितरण फलन के द्धारा इन प्रतिक्रियों के अनुपात को वास्तविक विचर (z) के स्केल(scale) पर बदला जाता है जोकि उपलब्ध परिवार माध्य वंशागतित्वके अनुमान की उपज करता है।

$$h^2_{f(\Delta P/beta)} = \frac{\Phi^{-1}\left(\bar{p}_{f/\delta}=1\right)-\Phi^{-1}\left(\bar{p}\right)}{\Phi^{-1}\left(\bar{p}_{pf/\delta}=1\right)-\Phi^{-1}\left(\bar{p}\right)} \qquad \text{...(23)}$$

सहायक विशेषकों के लिए स्टेएबिलिटीका समायोजन

लक्षणों की तरह स्टेएबिलिटी उत्पादन जैसे सहायक लक्षण तथा अन्य प्रकार के लक्षणों द्वारा सार्थक तरीके से प्रभावित होते हैं इसलिए स्टेएबिलिटी की वंशागति की सही तस्वीर पाने के लिए सहायक लक्षणों के प्रभाव को निरसन(eliminate) करने की सलाह दी जाती है। उदाहरण के तौर पर पशु डेरी के झुण्ड जीवन उत्तरजीविता(survival) तथा उत्पादन विशेषक के रूप में होती है, जोकि निम्नलिखित समीकरण द्धारा निर्देशित है।

$$P_{HL} = m_Y P_Y + m_S P_S \qquad \text{.....(24)}$$

जहाँ $p_{\mathrm{HL}}, p_y, p_s$क्रमशः झुण्ड जीवन के दृश्य प्रारूप मान, उत्पादन तथा उत्तरजीविता है। क्रमशः p_yऔर p_s पर p_{HL}के मानवीकृत आंशिक समाश्रयण गुणांक (standardized partial regression coefficient) m_Yऔर m_Sहैं। उत्पादन के लिए समायोजित झुण्ड जीवन के एक नये दृश्य प्रारूप विचर कुछ इस प्रकार आसानी से प्राप्त किये जा सकते हैं।

$$P_{HL/Y} = P_{HL} - r_{Y,HL}\, P_Y \quad = m_S(P_S - r_P\, P_Y) \qquad \text{.....(25)}$$

अब $p_{\mathrm{HL}/Y}$ को असली विचर लेते हुए, नये विचर को दी गयी सफलता की प्रायिकता के लिए रुड़न के विभिन्न बिन्दुओं की मदद से द्विपद विचर में बदला जाता है, उत्पादन के लिए समायोजित पशुओं के झुण्ड जीवन की वंशागतित्व का अनुमान आसानी से प्राप्त किया जा सकता है। अब समायोजित लक्षणों की वंशागतित्व का अनुमान वंशागति की सही तस्वीर प्रदर्शित करेगा जबकि पशुओं के झुण्ड जीवन की वास्तविक मान सहायक लक्षण उत्पादन द्धारा सार्थक तरीकों से प्रभावित हो सकते हैं।

आपेक्षित मूल माध्य वर्ग त्रुटि

विभिन्न विधियों की तुलना उसके परिशुद्धता की कुछ मापदण्डों के आधार पर की गयी है क्योंकि सभी अनुमान अनभिन्नत नहीं हैं। इसलिए प्रसरण के अनुमान सही अंदाजा नहीं दे सकते। अभिनत तथा कुछ परिशुद्धता के मापों के मान जानने के लिए, एक माप जोकि आपेक्षित रूट माध्य वर्ग त्रुटि के द्धारा कुछ इस प्रकार परिभाषित किया जाता है।

$$RMSE\,\% = \frac{\left[E\left(estimate - '\,true\ value'\right)^2\right]^{0.5}}{'\,true\ value'} \times 100 \qquad \text{.....(26)}$$

असंतुलितता की मात्रा

असंतुलितता की मात्रा को इस प्रकार परिभाषित कर सकते हैं

$\Delta = N(n-\lambda)$tgWk$n = N/S$, $\sum_{i=1}^{s} n_i = N$

$$\lambda = \frac{1}{S-1}\left[\sum_i n_i - \frac{\sum_i n_i^2}{N}\right]$$

यॅहा S = जनको या साड़ की संख्या

n_i =जनक (साड़) की पुत्री की संख्या

N =सम्पूर्ण पुत्रियों की संख्या

रॉा आँकड़ो पर वंशागतित्व का अनुमान

मोन्टे कारलो अनुकरण द्वारा जनित आंकड़े ऑफ–सिब मॉडल अनुसरण करता है।

$$Z_{ijk} = \mu + S_i + e_{ijk}$$

सही वंशागतित्व या रॉ आँकड़ों पर आधारित वंशागतित्व वह वंशागतित्व है जो कि द्विपद आँकड़ो या प्रारम्भिक लक्षणों को असली ऑफ–सिब अनुकरणीत आँकड़ों का प्रयोग करके संगणित की गयी है या की जाती हैं।
व्यक्तिगत संकीर्ण संवेदिये वंशागतित्व

$$\hat{h}^2_{(Z)} = \frac{4\hat{\sigma}^2_f(z)}{\hat{\sigma}^2_f(z) + \hat{\sigma}^2_e(z)} \qquad \text{....(27)}$$

अनुमानित घटक, प्रसरण के विश्लेषण (हन्डरसन विधि III,सरले इत्यादि 1992) उपरोक्त मॉडल को अनुप्रयुक्तकर प्राप्त किये जाते हैं।

सही पारिवारिक माध्य वंशागतित्व इस प्रकार है।

$$\hat{h}^2_{f(Z)} = \frac{\hat{\sigma}^2_f(z)}{\hat{\sigma}^2_f(z) + \hat{\sigma}^2_e(z)/n_{block} \times n_{plot}} \qquad \text{....(28)}$$

परिणाम और विचार–विमर्श

स्टेएबिलिटी की वंशागतित्व के अनुमान की दो विधियों की तुलना करने के लिए विभिन्न प्रकार के आँकड़े जिनकी विभिन्न प्रकार की असंतुलितता की मात्रायें है वह वंशागतित्वकी विभिन्न प्राचल (parameter) के लिए कम्प्यूटर से अनुकरणित किये गये हैं।
Z_{ijk}पर आँकड़े रेखीय मॉडल के अनुसार से जनित हैं।

$Z_{\text{ijk}} = \mu + S_i + e_{\text{ijk}}$एक सामान्यतः विस्तृत अनुपालनीय विचर Zके लिए जिसका पूर्ण (total) प्रसरण एक (1.0) है जोकि यादृच्छिक सम्पूर्ण ब्लॉकों में ऑफ–सिब की श्रेणी (series) में है।

परिवार मान (S_i) सामान्य प्रसरण की तरह अनुकरणीत है जिसका माध्य शून्य तथा प्रसरण 0.0125, 0. 025, 0.0375 और 0.0625 है। त्रुटियां मानिकी वातावरणीय मान(e_{ijk}) एकल गोसियन विचर के रूप में अनुकरणीत है। जिसका माध्य शून्य है तथा प्रसरण $(1 - \sigma^2_f)$है। रूडन के पांच बिन्दुओं या प्रारम्भिक लैवलस वास्तविक आंकड़ों को द्विपद आंकड़ों में बदलने के लिए प्रयोग किया जाता है। वह प्रारम्भिक जिनका प्रयोग किया गया है। p =0.05, 0.010, 0.15, 0.20, 0.25 जोकि द्विपद विशेषक को अनुपालन करने की प्रायिकतायें हैं। आँकड़ों की स्टेएबिलिटी (H^2_s =0.05, 0.010, 0.15, 0.20, 0.25) की वंशागतित्व के विभिन्न प्राचल का प्रयोग कर जनित किया जाता है। पाचलिता मान के लिए बीस जनकों के लिए नमूने जनित किये जाते हैं। जिनकी पुत्रियां पांच से चोबिस के बीच में होती हैं। इस प्रकार जनित अनुकरणित आंकड़ें स्टेएबिलिटी की वंशागतित्व के अनुमानों की विभिन्न प्रक्रियाओं के लिए अधीन है तथा इस प्रकार प्राप्त किये गये परिणाम तालिका–1 में दर्शित हैं तालिका–1 से यह

देखा जाता है कि संकीर्ण संवेदीय बीटा–द्विपद संपादित वंशागतित्व*($h^2_{rea(b)}$)* किसी भी दूसरे अनुमानों से अच्छा परिणाम देती है। डैमस्टर लरनर अनुमान लगभग सामान्यतः प्रभावशाली है लेकिन परिवार माध्य वंशागतित्व निहायति अभिन्न है जानने के लिए यह एक रोचक बिन्दू है कि असंतुलिता के कारण मानक त्रुटि अत्यधिक बढ़ जाती है।

तालिका–1: असामान परिवार साइज के मामलों में दिये हुयेh^2_S(स्टेएबिलिटी की वंशागतित्व) कें विभिन्न मानों के लिए पशुओं के झुण्ड जीवन का व्यक्तिगत संकीर्ण संवेदीय वंशागतित्व(h^2) और परिवार माध्य वंशागतित्व$\left(h^2_f\right)$ के औसत अनुमान।

अनुमान	h^2_S =0.05	h^2_S =0.10	h^2_S =0.15	h^2_S =0.20	h^2_S =0.25
h^2_Z	0.0502	0.1001	0.1503	0.2001	0.2450
	(0.0354)	(0.0525)	(0.0702)	(0.0848)	(0.0598)
$h^2_{rea(b)}$	0.0465	0.0987	0.1521	0.2092	0.2675
	(0.0671)	(0.0879)	(0.0879)	(0.1344)	(0.1600)
h^2_{DL}	0.0460	0.0977	0.1493	0.2045	0.2598
	(0.0660)	(0.0870)	(0.0961)	(0.1323)	(0.1570)
$h^2_{f(Z)}$	0.4105	0.6032	0.7031	0.7649	0.8086
	(0.2292)	(0.1560)	(0.1145)	(0.0905)	(0.0786)
$h^2_{f(beta)}$	0.1546	0.3310	0.4465	0.5306	0.5539
	(0.3098)	(0.2490)	(0.2085)	(0.1786)	(0.1554)
$h^2_{f(\Delta P)/beta)}$	0.1540	0.3295	0.4449	0.5206	0.5907
	(0.3084)	(0.2478)	(0.2056)	(0.2128)	(0.1547)

औसत मानक विचलन ब्रकिट में हैं

असंतुलितता की मात्रा= 35.0001.

उत्पादन के लिए समायोजन

प्राचलित(पैरामैट्रिक) मानh^2_Y का प्रयोग करके स्टेएबिलिटी के लिए आँकड़ें (उत्पादन की वंशागतित्व) =0.25,m_Y (उत्पादन पर पशुओं के झुण्ड जीवन के मानविकृत समाश्रयण गुणांक)=0.4, $r_{Y,HL}$ (पशुओं के झुण्ड जीवन और उत्पादन दृश्य प्रारूप सहसंबंध)=0.25, इन प्राचल मानों के लिए तर्क जिसके साथ–साथ स्टेएबिलिटी की विभिन्न

वंशागतित्व(h_S^2 =0.05,0.10,0.15,0.20,0.25) भी डैकर्स (1993) से लिया गया है। इस प्रकार प्राप्त समायोजित आँकड़ें आगे चल कर द्विपद स्कैल में बदल दिया जाता है। समायोजित स्टेएबिलिटी आँकड़ों के लिए परिणाम तालिका–2 में दिखाया गया है। डैमस्टर लरनर तथा संकीर्ण संवेदीय बीटा–द्विपद वंशागतित्व अनुमानों के मामलों में बेहतर परिणाम देखे गये हैं। बाकी दूसरे प्रक्रिया के लिए परिणाम अत्याधिक अभिनत है। यह देखना बहुत रोचक है कि समायोजिता के कारण ना सिर्फ अनुमान जनसंख्या प्राचल के बहुत नजदीक थे। बल्कि त्रुटियॉ अत्याधिक कम हो गयी और इस प्रकार अनुमानो की यथार्थता बढ़ा दी।

तालिका–2:असामान परिवार साइज के मामलों में दिये हुयेh_S^2(स्टेएबिलिटी की वंशागतित्व)के विभिन्न मानों के लिए उत्पादन के लिए समायोजित पशुओं के झुण्ड जीवन की व्यक्तिगत संकीर्ण संवेदीय वंशागतित्व(h^2)और परिवार माध्य वंशागतित्व$\left(h_f^2\right)$ के औसत अनुमान।

Estimate	h_S^2 =0.05	h_S^2 =0.10	h_S^2 =0.15	h_S^2 =0.20	h_S^2 =0.25
h_Z^2	0.0503	0.0977	0.1457	0.1937	0.2420
	(0.0334)	(0.0484)	(0.0615)	(0.0751)	(0.0884)
$h_{rea(b)}^2$	0.0524	0.1027	0.1536	0.2079	0.2628
	(0.0679)	(0.0872)	(0.1050)	(0.1278)	(0.1503)
h_{DL}^2	0.0534	0.1027	0.1529	0.2058	0.2586
	(0.0681)	(0.0868)	(0.1075)	(0.1320)	(0.1563)
$h_{f(Z)}^2$	0.5113	0.7056	0.7912	0.8393	0.8702
	(0.2951)	(0.1451)	(0.0929)	(0.0676)	(0.0528)
$h_{f(beta)}^2$	0.1866	0.3488	0.4550	0.5338	0.5926
	(0.3015)	(0.2449)	(0.2041)	(0.1745)	(0.1541)
$h_{f(\Delta P)/beta)}^2$	0.1857	0.3457	0.4529	0.5313	0.5899
	(0.3002)	(0.2482)	(0.2031)	(0.1733)	(0.1544)

औसत मानक विचलन ब्रकिट में हैं

असंतुलितता की मात्रा= 35.0001.

रूट माध्य वर्ग त्रुटि

विभिन्न विधियों की अनुभाविक तुलना के लिए औसत रूट माध्य वर्ग त्रुटि बहुत उपयोगी पायी गया है तथा विभिन्न वंशागतित्व और विभिन्न प्रारिम्भक प्रायिकताओं पर परिकल्पना की गयी है। बीस जनको जिसमें पाँच से चौबीस पुत्रियों और जो ब्लाक साईज़ पॉच में से उनके लिए रूडन प्रायिकताओं औसत की गई आपेक्षिक रूट माध्य वर्ग त्रुटि तालिका–3 में दिखायी गयी है और स्टेएबिलिटी की वंशागतित्वों के विभिन्न मानों पर समान औसत तालिका–4 में दिखायी गयी है तालिका–3 तथा तालिका–4 से यह साफतौर पर देखा गया है कि परिवार माध्य

वंशागतित्व अनुमानों के लिए आपेक्षित रूट माध्य वर्ग त्रुटि बाकी किसी भी वंशागतित्व अनुमान से संख्यात्मकता ज्यादा सार्थकता पूर्ण है। तालिका–4 से यह देखा गया है कि सही आँकड़ों बिन्दुओं पर परिवार माध्य वंशागतित्वों की आपेक्षिक रूट माध्य वर्ग त्रुटि उच्चतम मान है।

तालिका–3 असमान पुत्री के मामलों में पशुओ के झुण्ड जीवन की वंशागतित्व के चयनित अनुमानों की आपेक्षिक रूट माध्य वर्ग त्रुटि(RMSE%)

अनुमान	$h^2_{rea(b)}$	h^2_{DL}	$h^2_{f\,(beta)}$	$h^2_{f\,(\Delta P)/beta)}$	h^2_Z	$h^2_{f\,(Z)}$
h^2_S =0.05	134.8040	132.2291	670.1510	666.942	70.9687	854.3492
	(131.040)	(136.600)	(681.601)	(678.257)	(66.697)	(1055.292)
h^2_S =0.10	87.9778	85.5209	351.2075	349.3008	52.5299	526.8245
	(86.012)	(86.945)	(359.704)	(310.002)	(47.592)	(622.237)
h^2_S =0.15	72.7759	70.7717	248.2335	206.4008	46.8208	376.5623
	(70.074)	(71.692)	(250.656)	(249.152)	(41.091)	(431.909)
h^2_S =0.20	67.3535	66.1898	192.4388	199.7637	42.3125	286.0470
	(64.049)	(64.866)	(191.858)	(190.557)	(37.681)	(321.4421)
h^2_S =0.25	64.3965	62.9858	154.3059	152.6061	39.9093	224.4113
	(60.3965)	(62.637)	(150.892)	(150.893)	(35.516)	(248.963)

समायोजिताके मामलों में समरूप रूट माध्य वर्ग त्रुटि ब्रैकिट में है।

तालिका 4:– असमान पुत्रियों के मामले में पशुओ के झुण्ड जीवन में अनुमानों की आपेक्षिक रूट माध्य वर्ग त्रुटि(RMSE%)

Estimate	$h^2_{rea(b)}$	h^2_{DL}	$h^2_{f\,(beta)}$	$h^2_{f\,(\Delta P)/beta)}$
$\bar{p}$ =0.05	118.3930	117.7272	349.6306	348.0157
	(117.765)	(125.393)	(353.457)	(350.779)
$\bar{p}$ =0.10	91.169	89.4294	334.3121	332.5418
	(87.809)	(91.196)	(346.205)	(342.534)
$\bar{p}$ =0.15	72.4481	77.3592	329.7966	328.0628
	(76.218)	(75.920)	(335.362)	(336.125)
$\bar{p}$ =0.20	71.3920	68.1943	301.6859	299.0613
	(69.549)	(67.924)	(304.375)	(302.457)

$\bar{p}$ =0.25	67.6236	64.9872	300.3517	298.4655
	(65.008)	(62.305)	(299.289)	(296.969)

समायोजिता के मामलों में समरूप रूट माध्य वर्ग त्रुटि ब्रैकिट में है।

समायोजन के मामले में रूट माध्य वर्ग त्रुटि में सार्थक बदलाव नहीं देखा गया। स्टेएबिलिटी के अनुवांशिक मानों तथा रूडन बिन्दुओं में बढ़ते हुये चलन से आपेक्षिक बिन्दु के लिए वंशागतित्व के अनुमानों की आपेक्षिक रूट माध्य अनुमानों से अनुसरीत है। सही मानों तथा बीटा–द्विपद प्रक्रिया के परिणामों के अनुमानों तो कुछ एकरूपता दिखाते हैं। जबकि दूसरी प्रक्रिया में उत्पादन लक्षण पर आधारित आंकड़ों के समायोजन के लिए कोई विशेष ट्रेन्ड नहीं देखा गया। इन परिणामों से अन्ततः यह निष्कर्ष निकलता है अगर किसी को स्टेएबिलिटी केआँकड़े निर्धारित करने काप्रस्ताव हो तो इस निर्धारित आँकड़ों पर आधारित वंशागतित्व के अनुमान बहुत ही दक्षतापूर्ण तथा स्पष्ट अनुमान देगा और यदि किसी के पास स्टेएबिलिटी के सिर्फ द्विगुण आँकड़े हो तब बीटा–द्विपद अनुमान बाकी अनुमानों की विधियों की तुलना में बहुत ही सटीक और स्पष्ट हैं। जबकि असंतुलिता कई बार अनुमानों में बड़ी मानक त्रुटियों की वजह बन जाती है।

संदर्भ

1. डेक्कर्स जैक, सी.एम., 1993. थियोरेटिक्ल बेसिस फॉर जेनेटिक पैरामीटर ऑफ हर्ड लाईफ एण्ड इफेक्टस ऑन रिसपोंस टू सलेक्शन. जे. डेयरी सांईस, 76: 1433–1443.
2. डेम्पसटर, ई.आर. एण्ड लर्नर, आई.एम., 1950. हैरिटेबिल्टि ऑफ थ्रशोल्ड करेक्टर्स. जेनेट., 35: 212–236.
3. फालकोनर, डी.एस., 1981. इन्ट्रोडक्शन टू क्वानटीटेटिव जेनेटिक्स, सेकेन्ड एडीशन, लांगमैन, लंदन.
4. गियानोला, डी., 1979 हैरीटेबिल्टि आफ पोलीकोटोमस करेक्टर्स. जेनेट., 93: 1051–1055.
5. जॉनसन, एन.एल. एण्ड कोटज़ एस., 1970. कनटिनयूअस यूनीवेरीयेट डिस्ट्रीब्यूशन, 2. जॉन विले एण्ड सन्स, न्यूयार्क.
6. मैगनुसेन. एस. एण्ड क्रिमर, ऐ. 1995. दी बीटा–बायनोमियल मॉडल फार एस्टीमेटिंग आफ बाइनरी ट्रेटस. थीयोरेटि. एपीली. जैनेट., 91: 544–552.
7. सरले, एस.आर. कासीला. जी. एण्ड मैक्कुलाच, सी.ई., 1992. वेरियेन्स कम्पोनेंटस. जॉन विले एण्ड सन्स, न्यूयार्क.

स्थायित्व विश्लेषण

प्रकाश कुमार

भा.कृ.अनु.प.- भारतीय कृषि सांख्यिकी अनुसंधान संस्थान, नई दिल्ली-110012

prakash.kumar@icar.gov.in

1. परिचय

विभिन्न शोध कार्यक्रमों में यह काफी आम है, एक ही तरह के परीक्षण को अलग अलग स्थानों, मौसमों अथवा वर्षों में लगाया जाता है । इसमें अनुसंधान का मुख्य उद्देश्य किस्म / नस्ल की स्थायित्व को ज्ञात करते हैं । यह पाया गया है कि एक फसल की विभिन्न किस्मों की प्रभावशीलता को आम तौर पर एक स्थान से दुसरे स्थान एवं और एक मौसम से दुसरे मौसम में भिन्न होता है । एक एकल परीक्षण केवल एक ही स्थान या एक मौसम की जानकारी प्रस्तुत करता है । इस प्रकार एक वैध सुझाव प्राप्त करने के लिए आम तौर पर अलग–अलग स्थानों अथवा अलग–अलग समयों पर एक ही परीक्षण को पुनरावृत्त करतें हैं , ताकि हमें एक स्थान से दुसरे स्थान में पाए जाने वाले विविधता, समय के साथ बदलाव अथवा दोनों का प्रभाव पता चल पायें । इस दशा में आँक□ों की एक संयुक्त विश्लेषण के लिए पुनरावृत्त प्रयोग जो की एक उपयुक्त सांख्यिकीय प्रक्रिया है,का पालन करना होगा।

pooled analysis में, दो मुख्य बिंदुओं पर ध्यान देते हैं (i) किस्मों की औसत प्रतिक्रिया का अनुमान लगाने के लिए और (ii) किस्मों की निरंतरता का परीक्षण एक जगह से दुसरे जगह अथवा भिन्न समय में अर्थात स्थानों या वर्षों के साथ किस्मों के प्रभाव की पारस्परिक क्रिया (GE interaction) का परीक्षण । उच्च स्थायित्व वाले किस्मों आम तौर पर कम उपजाउ होते हैं इसी लिय GE interaction किसी भी किस्म / नस्ल की श्रष्ठता का प्रदर्शन में महत्वपूर्ण कठिनाई का कारण बनता जब किस्मों / नस्लों को वातावरण की एक श्रृंखला पर तुलना करते है। इसलिए हम फसल एवं जानवरों के विकास के लिए संतुलित तरीकों का उपयोग करतें हैं । इसके लिए कम GE interaction को पता लगाने की जरूरत होती है। यहाँ हमारा प्रारंभिक मूल्यांकन स्थायित्व जीनोटाइप की पहचान करनाहै।

2. जी. ई. इंटरेक्शन;**GE interaction)**के विश्लेषण के लिए रेखीय समाश्रयण मॉडल

माना की i^{th}जीनोटाइप जहाँ $;i = 1,2,\ldots,t$द्धएवं j^{th}पर्यावरण जहाँ $;j = 1,2,\ldots,s$ द्धपरY_{ij},औसत प्ररूपी मान है।

$$Y_{ij} = \mu + d_i + (1 + \beta_i)e_j + \delta_{ij} + \bar{\varepsilon}_{ij} \qquad (2.1)$$

जहाँ μ सामान्य माध्य है, d_i,i^{th}जीनोटाइपका प्रभाव है, e_j , j^{th}पर्यावरण का प्रभाव है, $(1+\beta_i)$, e_j पर Y_{ij} समाश्रयण है, δ_{ij},i^{th}जीनोटाइपएवं j^{th}पर्यावरण के लिए समाश्रयण से विचलन है, और$\bar{\varepsilon}_{ij}$यादृच्छिक त्रुटि है। जीण ईण इंटरेक्शन प्रभाव के लिए g_{ij} के निम्नलिखित संबंध है:

$g_{ij} = \beta_i e_j + \delta_{ij}$और(2.1)से

$$\sum_i d_i = \sum_j e_j = \sum_i \delta_{ij} = \sum_j \delta_{ij} = \sum_{i,j} \delta_{ij} = 0$$

μ, d_i, e_j and β_iका least square estimates (न्यूनतम वर्ग विधि अनुमान):

$\hat{\mu} = \bar{Y}_{..} = (\sum_{i,j} Y_{ij})/st$

$\hat{d}_i = \bar{Y}_{i.} - \bar{Y}_{..}$; $\{\hat{e}_j = \bar{Y}_{.j} - \bar{Y}_{..}$

$b_i = 1 + \hat{\beta}_i = \sum_j Y_{ij}\hat{e}_j / \sum_j \hat{e}_j^2$और$\hat{g}_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$

जहाँ, $\bar{Y}_{i.} = (\sum_j Y_{ij}/s)$, $\{\bar{Y}_{.j} = (\sum_i Y_{ij}/t)$

i^{th} जीनोटाइपका माध्य वर्ग विचलन ($s^2_{d_i}$) से दिखाया जा सकता है ।

$s^2_{d_i} = \sum_j \hat{\delta}^2_{ij}/(s-2)$, जहाँए$\sum_i \hat{\delta}^2_{ij} = \sum_{i,j} Y_{ij} - s\bar{Y}^2_{.i} - b_i^2 \sum_j e_j^2$

मॉडल (2ण1) पर आधारित विश्लेषण के रूप तालिका 1 में प्रस्तुत है ।

सारणी 1: GxE इंटरेक्शन के लिए प्रसरण के विश्लेषण

स्रोत	d.f.	S.S	M.S
जीनोटाइप (G)	(t-1)	$s\sum_i g_i^2$	MS1
पर्यावरण (E)	(s-1)	$t\sum_j e_j^2$	

GxEइंटरेक्शन	(t-1)(s-1)		
समाश्रयण के बीच विविधता	(t-1)	$\sum_i \beta_i^2 (\sum_j e_j^2)$	MS2
अवशिष्ट(Residual)	(t-1)(s-2)	$\sum_{i,j} \delta_{ij}^2$	MS3
औसत त्रुटि	s(t-1)(r-1)		$\overline{S}_e^2$

स्वतंत्रता की डिग्री दो घटकों में विभाजित है:

(i) (t-1) स्वतंत्रता की डिग्री के साथ प्रतीगमन वर्गों का योग के बीच विविधता

(ii) (t-1)(s-2) स्वतंत्रता की डिग्री साथ वर्गों के योग शेष

अलग–अलग लाइनों के लिए समाश्रयण के महत्व के परीक्षण के लिए निम्नलिखित विश्लेषण किया जा सकता है।

स्रोत	d.f.	MS
समाश्रयण	1	$(1+\beta_i^2)(\sum_j e_j^2)$
अवशिष्ट	(s-2)	$\sum_j \delta_{ij}^2/(s-2)$
कुल	(s-1)	$\sum_j (y_{ij} - y_{i.}/s)^2/(s-1)$

यहाँ हम परिकल्पना कि परीक्षण कर रहे हैं जहाँ (1 β_i) गैर शून्य अनुमानित है । प्रत्येक जीनोटाइप के लिए $\beta_i^2 \sum_j e_j^2$घटक भी निकाली और उसी अवशिष्ट के साथ तुलना की जा सकती है।

3. स्थायित्व के जैविक और कृषिशास्त्रीय अवधारणा

ब्रीडर के लक्ष्य और विचाराधीन चरित्र के लिए स्थायित्व के दो अवधारणा है, जैविक और कृषिशास्त्रीय अवधारणा (Becker, 1981) है । इसे क्रमशः स्थिर और गतिशील अवधारणा (Leon, 1985)भी कहते हैं । जैविक अवधारणा के तहत जब जीनोटाइप का परीक्षण एक संख्या से अधिक वातावरण में करते हएक स्थिर जीनोटाइप जिसका फेनोटाइप Y_{ij} है, जो की चरित्र के स्तर $\overline{Y}_{i.}$ से कम विचलन दिखता हैं । अलग अलग वातावरण में स्थिर प्रदर्शन के करण, इस अवधारणा को स्थायित्व के अवधारणा भी कहा जाता है ।

कृषिशास्त्रीय अवधारणा में अनाज की उपज को देखते हैं, एक स्थिर जीनोटाइप अनुकूल वातावरण के तहत इतनी अच्छी पैदावार नहीं देता जितना की प्रतिकूल वातावरण में भी औसतन अच्छी पैदावार करता है । ब्रीडर एक जीनोटाइप जो प्रत्येक वातावरण में एक जैसा प्रतिक्रिया दे अर्थात ऐसा किस्म जो जीण्ई्ण इंटरेक्शन ;*i.e.* $(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..}) \simeq 0)$, नहीं दिखाता है, पसंद करता है ।व्यापक रूप से इस्तेमाल पैरामीट्रिक स्थायित्व उपायों और अपनी अंतर्निहित स्थायित्व अवधारणा का विहंगावलोकन तालिका 2 में दिखाया गया है।

सारणी 2: आम स्थायित्व मापों और उनके अंतर्निहित स्थायित्व अवधारणा

स्थायित्व माप	प्रतीक	अंतर्निहित स्थायित्व अवधारणा
Environmental variance	$S^2_{Y_i}$	जैविक
Ecovalence	W_i	कृषिशास्त्रीय
Stability variance	$\hat{\sigma}^2_i$	कृषिशास्त्रीय
Regression coefficient	b_i	जैविक/कृषिशास्त्रीय
Deviation mean square	$s^2_{d_i}$	कृषिशास्त्रीय
Coefficient of determination	r^2_i	कृषिशास्त्रीय
Hanson's stability measure	$\hat{D}^2_{(i.)}$	कृषिशास्त्रीय

प्रचलन में उपज स्थायित्व अधिकांश कृषिशास्त्रीय अवधारित है । जैविक अवधारणा के लिए केवल दो माप उपलब्ध हैं और वे environmental variance $S^2_{Y_i}$ and the environmental coefficient of variation (CV_i) है ।

$$S^2_{Y_i} = \sum_j (Y_{ij} - \bar{Y}_{i.})^2 / (s - 1) \qquad (3.1)$$

$$CV_i = (S_{Y_i} / \bar{Y}_{i.}) \times 100 \qquad (3.2)$$

जिनमें से कम मान हो हमेशा उच्च स्थायित्व दिखता है। हालांकि जो मापांक सैद्धांतिक रूप से काफी अच्छा हो निम्लिखित कारण से व्यावहारिक उपयोगिता नहीं होता है (i) जैविक अवधारणा के तहत स्थायित्व आमतौर पर अपेक्षाकृत कम उपज दिखता है (ii) उच्च स्तरिय उपज वाले जीनोटाइप को वृहत वातावरण की सीमा में अमल में लाना मुश्किल है।

4. रिक्की एकोवालेंस ;**Wricke's ecovalence measure**द्ध

जीनोटाइप के इंटरेक्शन रेसिडूअल के वर्गों के योग सेसरल और गणना में आसान, एकोवालेंस मापांक (W_i)मिलता है:

$$W_i = \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 = \sum_j \hat{g}_{ij}^2 \qquad (3.3)$$

क्म ecovalence मान वाले जीनोटाइप, उपज–स्थायित्व की दृष्टि से आदर्श माना जाता है ।

5. शुक्ला की स्थायित्व प्रसरण ;**Shukla's stability variance measure)**

रेसिडूअल$g_{ij} + \bar{\varepsilon}_{ij}$के प्रसरण का अनुमान $\hat{\sigma}_i^2$, i^{th}जीनोटाइप की स्थायित्व का एक उपयोगी सूचक होता है।$\hat{\sigma}_i^2$को स्थायित्व प्रसरण के रूप में परिभाषित किया गया है:

$$\hat{\sigma}_i^2 = \frac{t}{(s-1)(t-2)} W_i - \frac{MS(GE)}{(t-2)} \qquad (3.4)$$

जहाँW_i, (3.3) परिभाषित किया गया है और MS(GE) जी ई इंटरेक्शन के माध्य वर्ग है, [MS(GE) = $\sum_{i,j} \hat{g}_{ij}^2/(s-1)(t-1)$]. यह स्टेटिसटिक W_iका रैखिक संयोजन है इसीलिए रेंकिंग के लिए दोनों मापांक W_iऔर$\hat{\sigma}_i^2$बराबर है ।

6. हेनसन के स्थायित्व मापांक ;**Hanson's stability measure)**

हेनसन के जीनोटाइपिक स्थायित्व मापांक को $\hat{D}_{(i.)}^2$से परिभाषित करते हैं ।

$$\hat{D}_{(i.)}^2 = \sum [Y_{ij} - \bar{Y}_{i.} - b_{min}(\bar{Y}_{.j} - \bar{Y}_{..})]^2$$

$$= \sum [Y_{ij} - \bar{Y}_{i.} - b_{min} e_j]^2 \qquad (3.5)$$

जहाँ Eberhart और रसेल के अर्थ में b_{min},b_i $(i = 1,2,\ldots,t)$ का न्यूनतम मान है । इससे पता चलता है कि स्थिर जीनोटाइप वह है जो सीधे लाइन से विचलित नहीं होता है ।

$$Y_{ij} = \bar{Y}_{i.} + b_{min}(\bar{Y}_{.j} - \bar{Y}_{..})$$

7. निर्धारण गुणांक**(Coefficient of determination measure)**

यह स्थायित्व मापांक Pinthus (1973) द्वारा प्रस्तावित और r_i^2 के रूप में परिभाषित किया गया है । $r_i^2 = \frac{b_i^2 \sum_j e_j^2}{b_i^2 \sum_j e_j^2 + \sum_j \delta_{ij}^2}$ (3.6)

जहाँ $\sum_j e_j^2 = \sum(\bar{Y}_{.j.} - \bar{Y}_{...})^2$ और

$$\sum_j \delta_{ij}^2 = \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 - (b_i - 1)^2 \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2$$

$$= W_i - (b_i - 1)^2 \sum_j e_j^2$$

$(s-1)s_{d_i}^2$, समाश्रयण से वर्ग विचलन का योग को दर्शाता है । मापांक b_i, r_i^2माप के पैमाने से स्वतंत्र है। जीनोटाइप की स्थायित्व की रैंकिंग के लिएr_i^2के उच्च मान वांछित होता है ।

8. एबेर्हार्ट और रसेल के दो पैरामीटर**(Eberhart and Russell's two-parameter measure)**

एबेर्हार्ट और रसेल;**1966**द्ध समाश्रयण गुणांक b_i,को पहला स्थायित्व मापांक लिया है।

$$b_i = \sum_j (Y_{ij} - \bar{Y}_{i.})(\bar{Y}_{.j} - \bar{Y}_{..}) / \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 = 1 + [\sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..}) / \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2] \quad (3.7)$$

और दूसरा मापांक, $s_{d_i} = [\sum_j \delta_{ij}^2 (s-2)] - \bar{S}_e^2$ (3.8)

जहाँ$\bar{S}_e^2$, औसत त्रुटि है, $\bar{S}_e^2 = (\sum_j s_j^2 / \text{sr})$,जहाँ$s_j^2$विभिन्न प्रयोगों के लिए त्रुटि माध्य वर्ग है, जिसमें प्रत्येक, r रेप्लिकेसन के साथ संचालित है ।यहाँ $s_{d_i} = 0$ and $b_i = 1$ को स्थायित्व जीनोटाइप होता है ।

9. पर्किन्स और जिंक्स के दो पैरामीटर मापांक**(Perkins and Jinks two-parameter measure)**

Eberhart और रसेल मॉडल पर आधारित समाश्रयण तकनीक में मामूली संशोधन करके पर्किन्स और जिंक्स (1968)निम्नलिखित मापांक दिया है ।

$$\hat{\beta}_i = \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})(\bar{Y}_{.j} - \bar{Y}_{..}) / \sum_j (\bar{Y}_{.j} - \bar{Y}_{..})^2 \quad (3.9)$$

$$s_{d_i}^2 = \sum_j \delta_{ij}^2 (s-2) \quad (3.10)$$

जहाँβ_i, b_iसे सम्बंधित है, अर्थात् $\beta_i = b_i - 1$.

उदाहरणरू

वातावरण1:

	R1	R2	R3	R4
V1	1250	1150	1175	1000
V2	1350	1450	1150	1050
V3	1250	750	925	1100
V4	1300	1500	1425	1575
V5	1550	1375	1475	1300
V6	1300	1325	1525	1675
V7	1350	900	1475	1675
V8	1175	1500	1250	1225
V9	1150	1225	1525	1275
V10	1200	1250	900	1000
V11	1725	1725	2100	1650
V12	1500	1450	1125	1625
V13	1300	1000	1250	950
V14	1175	1200	1275	1400

वातावरण2:

	R1	R2	R3	R4
V1	1850	1650	1700	1150
V2	1800	1600	1800	1200
V3	2000	1950	1900	1700
V4	2200	2000	1800	2100
V5	1350	1650	1300	1250
V6	1900	1900	1550	1700
V7	1500	1500	1800	1400

V8	1000	1200	1750	1150
V9	1600	1900	1600	1700
V10	2100	1550	1650	2100
V11	1550	1400	1350	1550
V12	1500	1800	2050	2150
V13	900	1150	1050	900
V14	1900	1900	1700	1800

वातावरण3:

R1	R2	R3	R4
V1	1150	1210	1170	1200
V2	910	1200	1130	1100
V3	1030	1200	1100	850
V4	750	900	770	660
V5	690	620	800	675
V6	540	860	630	550
V7	1050	1070	1000	1260
V8	560	650	470	640
V9	830	1160	1170	1100
V10	1270	900	900	830
V11	400	950	780	1010
V12	1130	1450	1150	1250
V13	1450	1500	1200	1350
V14	600	850	660	760

वातावरण4:

	R1	R2	R3	R4
V1	880	860	895	910
V2	870	880	785	870
V3	755	750	795	760
V4	770	810	810	935
V5	890	945	900	915
V6	800	800	710	870
V7	1095	995	1050	1045
V8	1000	980	890	980
V9	920	895	955	990
V10	885	930	805	855
V11	1045	900	1045	1065
V12	860	805	890	955
V13	860	975	940	970
V14	980	890	805	865

10. परिणाम और चर्चा

उदाहरण 1 केलिए स्थिरतामापक आधारितजीनोटाइप / किस्मोंकीरैंकिंग

किस्में	माध्य	b_i	$\sum_j \hat{\delta}_{ij}^2$	CV_I	W_i	r_i^2	$\hat{\sigma}_i^2$
V1	6	3	8	3	5	10	5
V2	7	5	4	4	2	6	2
V3	10	11	13	12	11	9	11
V4	2	14	3	14	12	2	12
V5	13	6	10	7	7	11	7

V6	11	13	6	13	9	4	9
V7	4	2	1	2	3	1	3
V8	14	4	12	8	10	12	10
V9	5	7	2	5	1	3	1
V10	8	4	9	10	8	8	8
V11	3	10	14	9	13	13	13
V12	1	9	7	6	4	7	4
V13	12	1	11	1	14	14	14
V14	9	12	5	11	6	5	6

स्थायित्व विश्लेषण–अप्राचालिक दृष्टिकोण

फसल सुधार की गति विधियों की सफलता काफी हद तक बड़े पैमाने पर उत्पादन के लिए बेहतर किस्मों की पहचान पर निर्भर करता है। एक जीनोटाइप बेहतर माना जा सकता है अगर यह अनुकूल वातावरण के तहत उच्च उपज क्षमता और एक ही समय में प्ररूपी स्थायित्व है। फेनोटाइप, जीनोटाइप(G), पर्यावरण (E) एवं जी. ई. इंटरेक्शन (GEI) का एक मिश्रण है । स्थिरता की अवधारणा को समझने के लिए, शोधकर्ताओं इन पदों जैसे अनुकूलन, प्ररूपी स्थायित्व और उपज स्थायित्व का उपयोग करतें हैं (Becker and Leon, 1988) ।

यहाँ पद अनुकूलन के बारे में दावोलकर (1999) ने विस्तृत चर्चा की है, सभी जीवित चीजों में शारीरिक समायोजन होता है जो उन्हें अपने तात्कालिक परिवेश में उतार–चढ़ाव से सामना करने की अनुमति देती है । ये समायोजन खुद को अनुकूलन के रूप में जाना जाता है । अनुकूलन एक जीनोटाइप का लक्षण जो चयन के तहत अपने अस्तित्व के लिए परमिट करती है । सायमंड्स (1981) ने अनुकूलन दो पहलू दिया है ।

1-विशिष्ट जिनोटायपिक अनुकूलनरू यह एक सीमित पर्यावरण के लिए अनुकूलन के करीब है ।

2-सामान्य जिनोटायपिक अनुकूलनरू यह एक विभिन्न वातावरण के लिए अनुकूलन है ।

ब्रीडरध्द किसान के अंतिम लक्ष्य और विचाराधीन विशे”ाता के लिए स्थायित्व के दो अवधारणा है, जैविक और कृषिशास्त्रीय अवधारणा है ।इसे क्रमशः स्थिर और गतिशील अवधारणा (Leon, 1985)भी कहते हैं ।

जैविक अवधारणा के तहत एक स्थिर जीनोटाइप जिसका फेनोटाइप Y_{ij} है, जो की विशे"ाता के स्तर $\overline{Y}_{i.}$ से कम विचलन दिखता है जब जीनोटाइप का परीक्षण एक संख्या से अधिक वातावरण में करते हैं । अलग अलग वातावरण में स्थिर प्रदर्शन के करण, इस अवधारणा को स्थायित्व के अवधारणा भी कहा जाता है ।

जैविक अवधारणा में अनाज की उपज को देखते हैं, एक स्थिर जीनोटाइप अनुकूल वातावरण के तहत इतनी अच्छी पैदावार नहीं देता जितना की प्रतिकूल वातावरण में भी औसतन अच्छी पैदावार करता है ।

सारणी 2: स्थायित्व मापों और उनके अंतर्निहित स्थायित्व अवधारणा

स्थायित्व माप	प्रतीक	अंतर्निहित स्थायित्व अवधारणा
Environmental variance	$S^2_{Y_i}$	जैविक
Ecovalence	W_i	कृषिशास्त्रीय
Stability variance	$\hat{\sigma}^2_i$	कृषिशास्त्रीय
Regression coefficient	b_i	जैविक/कृषिशास्त्रीय
Deviation mean square	$s^2_{d_i}$	कृषिशास्त्रीय
Coefficient of determination	r_i^2	कृषिशास्त्रीय
Hanson's stability Measure	$\hat{D}^2_{(i.)}$	कृषिशास्त्रीय
Huhn's Measures	$S_i^{(2)}$, $S_i^{(3)}$, $S_i^{(4)}$, $S_i^{(5)}$, $S_i^{(6)}$	जैविक/कृषिशास्त्रीय
Kang'srank-sum	RS	कृषिशास्त्रीय
Ketata's ranking sum	KRS	कृषिशास्त्रीय

प्रचलन में उपज स्थायित्व अधिकांश कृषिशास्त्रीय अवधारित है । जैविक अवधारणा के लिए केवल दो माप उपलब्ध हैं और वे environmental variance $S^2_{Y_i}$ and the environmental coefficient of variation (CV_i) है ।

$$S^2_{Y_i} = \sum_j (Y_{ij} - \bar{Y}_{i.})^2 / (s-1) \qquad (3.1)$$

$CV_i=(S_{Y_i}/\bar{Y}_{i.})$ x 100 (3.2)

जिनमें से कम मान हो हमेशा उच्च स्थायित्व दिखता है। हालांकि जो मापांक सैद्धांतिक रूप से काफी अच्छा हो निम्लिखित कारण से व्यावहारिक उपयोगिता नहीं होता है (i) जैविक अवधारणा के तहत स्थायित्व आमतौर पर अपेक्षाकृत कम उपज दिखता है (ii) उच्च स्तरिय उपज वाले जीनोटाइप को वृहत वातावरण की सीमा में अमल में लाना मुश्किल है।

स्थायित्व विश्लेषण को दो तरीकें, प्राचालिक और अप्राचालिक समूहों में वर्गीकृत करते हैं (Huhn, 1979)। प्राचालिक दृष्टिकोण जीनोटाइप, पर्यावरण और GEIप्रभाव के वितरण के बारे में सांख्यिकीय मान्यताओं, पर आधारित है । लेकिन, अप्राचालिक कोई विशेष सांख्यिकीय मान्यताओं, पर आधारित नही है ।

अप्राचालिक स्थायित्व विधि, प्राचालिक स्थायित्व विधि से कई मामले में अच्छा है, पर्यवेक्षित मान के वितरण के बारे में कोई मान्यताओं की जरूरत नहीं है, यह outliers से उत्पन्न पक्षपात को कम करता है एवं उपयोग तथा व्याख्या करने के लिए आसान है, और एक या अधिक जीनोटाइप को हटाने एवं जोड़ने से परिणामां में ज्यादा भिन्नता नहीं देता है ।

सच है कि अप्राचालिक विधि, प्राचालिक विधि समकक्ष से कम शक्तिशाली है, लेकिन जब जीनोटाइप की संख्या अधिक हो तो अप्राचालिक विधि की क्षमता प्राचालिक विधि के क्षमता के काफी समकक्ष होती है (रिगर और प्रभाकरण, 2000)। स्थायित्व मापने के लिए बहुत सारे विधियाँ है । प्रजनन कार्यक्रमों/किस्म के चयन में जीनोटाइप परीक्षण के लिए कुछ estimators को जोड़ने से अच्छा और उचित estimators के लिए अप्राचालिक विधि एवं प्राचालिक विधि के बीच सांख्यिकीय संबंधों का अध्ययन करने के लिए आवश्यक है।

2.सामग्री और तरीके

इस अध्ययन में इस्तेमाल डेटा आंध्र प्रदेश के विभिन्न कृषि जलवायु क्षेत्रों में स्थित अनुसंधान केंद्रों पर आयोजित की बहु स्थान साल परीक्षणों से एकत्र किए गए थे। प्रायोगिक लेआउट 3 रेप्लिकेसन के साथ यादृच्छिक ब्लॉक डिजाइन में किया गया था। डेटा क्षेत्रीय कृषि अनुसंधान स्टेशन (RARs), Palem, ANGRAU, आंध्र प्रदेश के द्वारा आपूर्ति की गई है । फली पैदावार प्रतिहेक्टेयर किलो के रूप में व्यक्त किया गया। 10 जीनोटाइप और 12 वातावरण के माध्य डेटा (तालिका 1) में प्रस्तुत है ।

तालिका 1:10 जीनोटाइप एवं 12 वातावरण के माध्य उपज डेटा

	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	E11	E12
G1	1773	880	2841	2020	856	1382	1458	282	1190	1001	2708	1832
G2	1715	861	2497	2020	505	1104	1153	275	1394	882	1956	1907
G3	1241	424	3266	1717	1148	1225	1130	113	701	705	1688	1568
G4	1472	917	3172	2222	1505	1475	1222	632	1308	334	2833	1157
G5	1208	1435	3625	1919	903	1432	921	862	1081	539	2303	1778
G6	1893	1310	2716	2374	1320	1476	1482	680	1498	591	2877	2333
G7	1852	1169	2527	2222	903	1220	1407	455	1637	521	2042	1732
G8	1266	993	2245	1869	292	972	1171	275	1419	767	2184	2037
G9	1736	792	2376	2172	981	1113	1051	364	1579	364	2940	1500
G10	1442	695	2800	2071	1051	1890	1051	605	1684	67	2083	1419

3. Nassar and Huhn (1987)के द्वारा जीनोटाइप के रैंक के आधार पर प्ररूपी स्थायित्व के कुछ अप्राचालिक विधियों का विवरण में नीचे दिए गए हैंरू

3.1). प्रत्येक वातावरण में प्रसरण रैंक**($S_i^{(2)}$)**

$S_i^{(2)} = \frac{\sum_{j=1}^{E}(r_{ij} - \bar{r}_{i.})^2}{(E-1)}$जहाँ E वातावरण की संख्या है,

$\bar{r}_{i.}$ वातावरण में रैंक का माध्य है, r_{ij} प्रत्येक वातवरण में इंटरेक्शन रेसिडूअल V_{ij} पर आधारित जीनोटाइप का रैंक है, जहाँ रैंक कम से उच्चतम की ओर आवंटित है ।

3.2). $\bar{r}_{i.}$के प्रति इकाई में जीनोटाइप के निरपेक्ष रैंक के अंतर का योग **($S_i^{(3)}$)**

$S_i^{(3)} = \frac{\sum_{j=1}^{E} |r_{ij} - \bar{r}_{i.}|}{\bar{r}_{i.}}$ जहाँ E वातावरण की संख्या है,$\bar{r}_{i.}$वातावरण में रैंक का माध्य है, r_{ij} प्रत्येक वातवरण में इंटरेक्शन रेसिडूअल V_{ij} पर आधारित i^{th} जीनोटाइप एवं j^{th} वातवरण का रैंक है, जहाँ रैंक कम से उच्चतम की ओर आवंटित है ।

3.3). प्रत्येक वातावरण में रैंक में मानक विचलन **($S_i^{(4)}$)**

$S_i^{(4)} = \sqrt{\frac{\sum_{j=1}^{E} (r_{ij} - \bar{r}_{i.})^2}{E}}$जहाँ E वातावरण की संख्या है,$\bar{r}_{i.}$वातावरण में रैंक का माध्य है, r_{ij} प्रत्येक वातवरण में इंटरेक्शन रेसिडूअल V_{ij} पर आधारित i^{th} जीनोटाइप एवं j^{th} वातवरण का रैंक है, जहाँ रैंक कम से उच्चतम की ओर आवंटित है ।

3.4). प्रत्येक वातावरण में रैंक मेंमाध्य विचलन**($S_i^{(5)}$)**

$S_i^{(5)} = \frac{\sum_{j=1}^{E} |r_{ij} - \bar{r}_{i.}|}{E}$ जहाँ E वातावरण की संख्या है,$\bar{r}_{i.}$वातावरण में रैंक का माध्य है, r_{ij} उपज के माध्यपर आधारित i^{th}जीनोटाइप एवं j^{th} वातवरण का रैंक है।

3.5). $\bar{r}_{i.}$के प्रति इकाई में प्रसरण रैंक**($S_i^{(6)}$)**

$S_i^{(6)} = \frac{\sum_{j=1}^{E} (r_{ij} - \bar{r}_{i.})^2}{\bar{r}_{i.}}$ जहाँ E वातावरण की संख्या है,$\bar{r}_{i.}$वातावरण में रैंक का माध्य है, r_{ij} प्रत्येक वातवरण में इंटरेक्शन रेसिडूअल V_{ij} पर आधारित i^{th}जीनोटाइप एवं j^{th} वातवरण का रैंक है, जहाँ रैंक कम से उच्चतम की ओर आवंटित है ।

4. रिक्की एकोवालेंस ;**Wricke's ecovalence measure**द्ध

इंटरेक्शन वर्गों के योग के लिए जीनोटाइप के योगदान सेसरल और गणना करने के लिए आसान, ,एकोवालेंस मापांक (W_i)मिलता है:

$$W_i = \sum_j (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})^2 = \sum_j \hat{g}_{ij}^2 \qquad (3.3)$$

कमecovalence मान वाले जीनोटाइप, उपज–स्थायित्व की दृष्टि से आदर्श माना जाता है ।

5. शुक्ला की स्थायित्व प्रसरण;**Shukla's stability variance measure)**

रेसिडूअल$g_{ij} + \bar{\varepsilon}_{ij}$के विचरण का अनुमान $\hat{\sigma}_i^2$, i^{th}जीनोटाइप की स्थायित्व का एक उपयोगी सूचक होता है।$\hat{\sigma}_i^2$को स्थायित्व विचरण के रूप में परिभाषित किया गया है:

$$\hat{\sigma}_i^2 = \frac{t}{(s-1)(t-2)} W_i - \frac{\text{MS(GE)}}{(t-2)} \qquad (3.4)$$

जहाँW_i, (3.3) परिभाषित किया गया है और MS(GE) जी ई इंटरेक्शन के माध्य वर्ग है, [MS(GE) = $\sum_{i,j} \hat{g}_{ij}^2/(s-1)(t-1)$]. यह स्टेटिसटिक W_iका रैखिक संयोजन है इसीलिए रेंकिंग के लिए दोनों मापांक W_iऔर$\hat{\sigma}_i^2$बराबर है ।

6. कांग रैंक-योग**(KangRank-Sum(RS))**

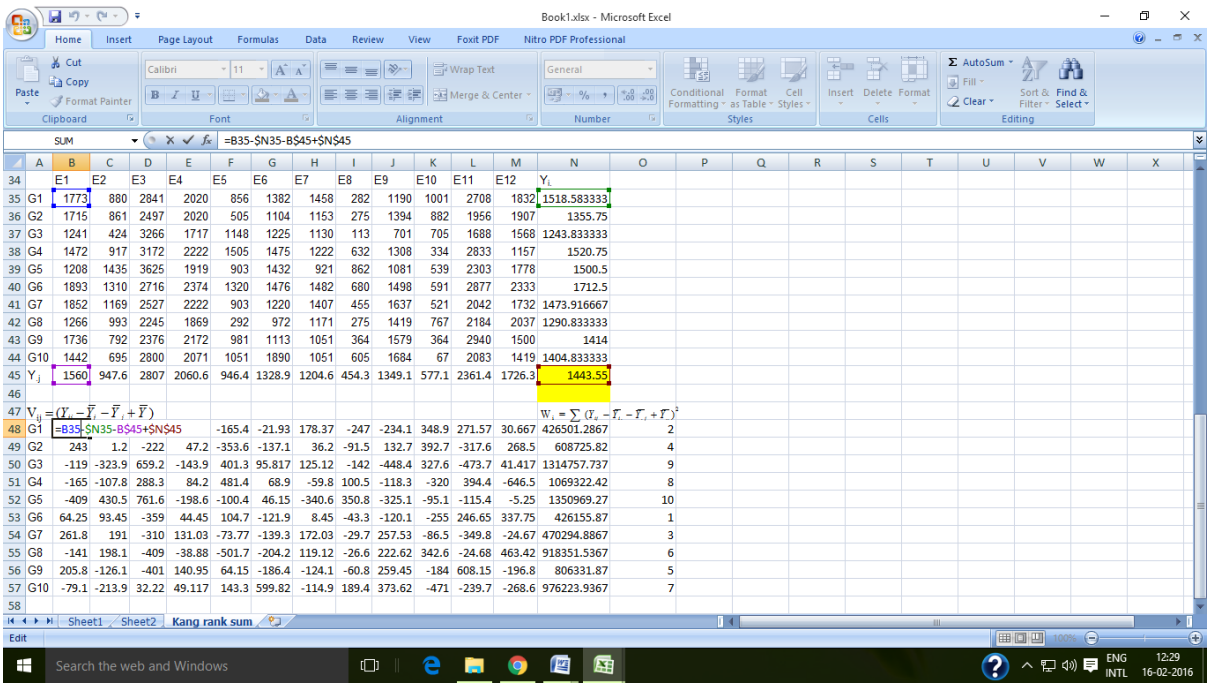
कांग रैंक–योग एक अप्राचालिक विधि है, जहां उपज एवं शुक्ला की प्रसरण दोनों को लिया जाता है। यहाँ उपज एवं स्थायित्व दोनों के लिए एक ही भार देतें हैं ताकि हमें उच्च उपज के साथ साथ स्थायित्व जीनोटाइप भी मिल सके । इस विधि में, उच्चतम उपज जीनोटाइप और सबसे कम स्थायित्व प्रसरण वाले जीनोटाइप दोनों को रैंक 1 देतें हैं उसके बाद सभी जीनोटाइप के लिए दोनों मानकों के रैंकों पर आधारित योग निकलते हैं । (Akcura& Kaya, 2008)के अनुसार प्रत्येक जीनोटाइप से कम रैंक वाले रैंक–योग (RS)सबसे अधिक वांछनीय जीनोटाइप माना जाता है ।

RS_i=माध्य उपज की रैंक + σ_i^2की रैंक

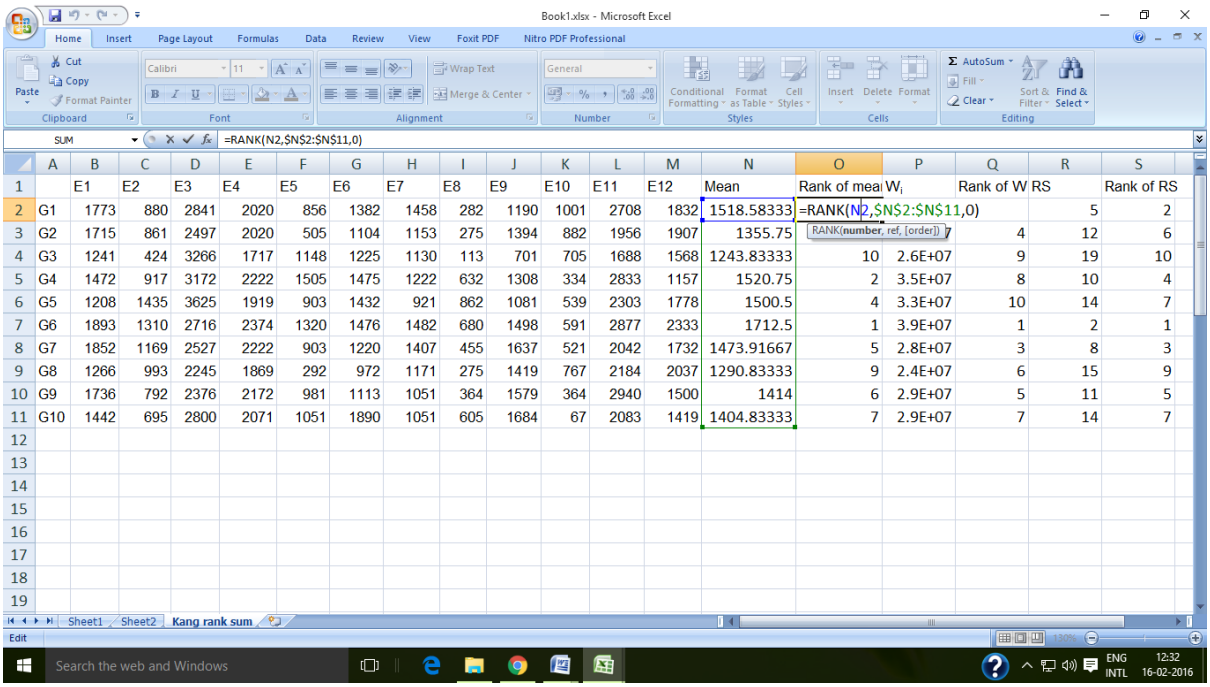
यहाँ रैंकिंग के लिए हम शुक्ला की स्थायित्व प्रसरण पैरामीटर (σ_i^2) कोWricke ecovalence(W_i)के सामान मानते हैं क्योंकि शुक्ला की स्थायित्व प्रसरण पैरामीटर(σ_i^2), Wricke ecovalence(W_i)का रेखीय संयोजन से बना होता है ।

□□□□□□रू.

सबसे पहले हम ऊपर लिए गए डेटा से MS Excel में $V_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$का मान निकलतें हैंए फिर Wricke ecovalence(W_i) का मान निकलते हैं, उसके बाद प्रत्येक जीनोटाइप का Wricke ecovalence(W_i) पर आधारित रैंक निकलते हैं जिसमेअधिकतमउपज–माध्य मान वाले जीनोटाइप को अधिकतम रैंक देते हैं ।



प्रत्येक जीनोटाइप का उपज के माध्य मान पर आधारित रैंक नकालते हैं जिसमे अधिकतम उपज–माध्य मान वाले जीनोटाइप को न्यूनतम रैंक देते हैं, उसके बाद दोनों मानकों से निकले गये रैंको का रैंकयोग निकलते हैं ।



निकले गये रैंक–योग(RS) का उपयोग वांछनीय जीनोटाइप का चयन में करते हैं ।

निष्कर्षः प्रत्येक जीनोटाइप में से सबसे कम रैंक वाले रैंक–योग (RS) सबसे अधिक वांछनीय जीनोटाइप माना जाता है ।

7. Ketata का रैंकिंग योग विधि (Ketata's ranking sum methods)

Ketata (1988) ने एक रैंकिंग विधि दिया है जिसमे जीनोटाइप का रैंक सभी प्रकार के वातावरण में अनाज की उपज पर आधारित होता है । प्रत्येक जीनोटाइप का रैंक उपज के माध्य एवं मानक विचलन पर आधारित होता है । इस विधि में जो जीनोटाइप अधिकतम प्रदर्शन करता है उसका रैंक 1 लेते हैं और अगर एक जीनोटाइप का प्रदर्शन माध्य रैंक 1 के करीब एवं कम रैंक के मानक विचलन के करीब हो सबसे अधिक स्थायी किस्म के रूप में जाना जाता है ।

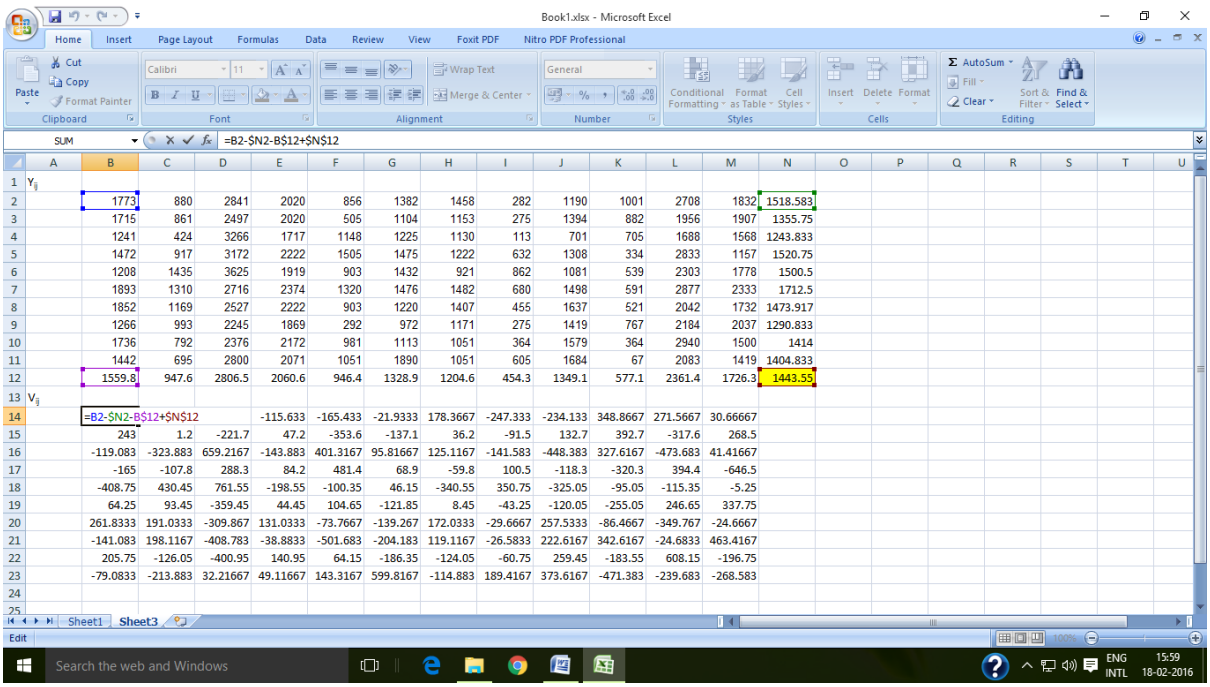
8. संशोधित स्थायित्व मापंक

यहाँ स्थायित्व मापंक निकलने के लिए दो मापांकों का उपयोग करते हैं पहला एक जीनोटाइप का सभी वातावरण से रैंक–योग जो की इंटरेक्शन रेसिडूअल $\hat{V}_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ or $(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)$ पर आधारित हो एवं दूसरा प्रत्येक वातावरणमें उसी जीनोटाइप प्रसरण जो की इंटरेक्शन रेसिडूअल $\hat{V}_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ or $(Y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)$ पर आधारित हो दोनों को लेते हैं । एक बार पुनः रैंक–योग का रैंक जिसमे अधिकतम मान वाले जीनोटाइप को अधिकतम रैंक देते हैं एवं इसी प्रकार जीनोटाइप प्रसरण के लिए भी रैंकिंग करते है जिसमे अधिकतम मान वाले जीनोटाइप को अधिकतम रैंक देतेहैं।दोनों मापांक का रैखिक संयोजन करके संशोधित स्थायित्व मापंक विकसित करते हैं जिसे रैंक आधारित स्टेबिलिटी इंडेक्स (RSI) से चिन्हित करते हैं ।

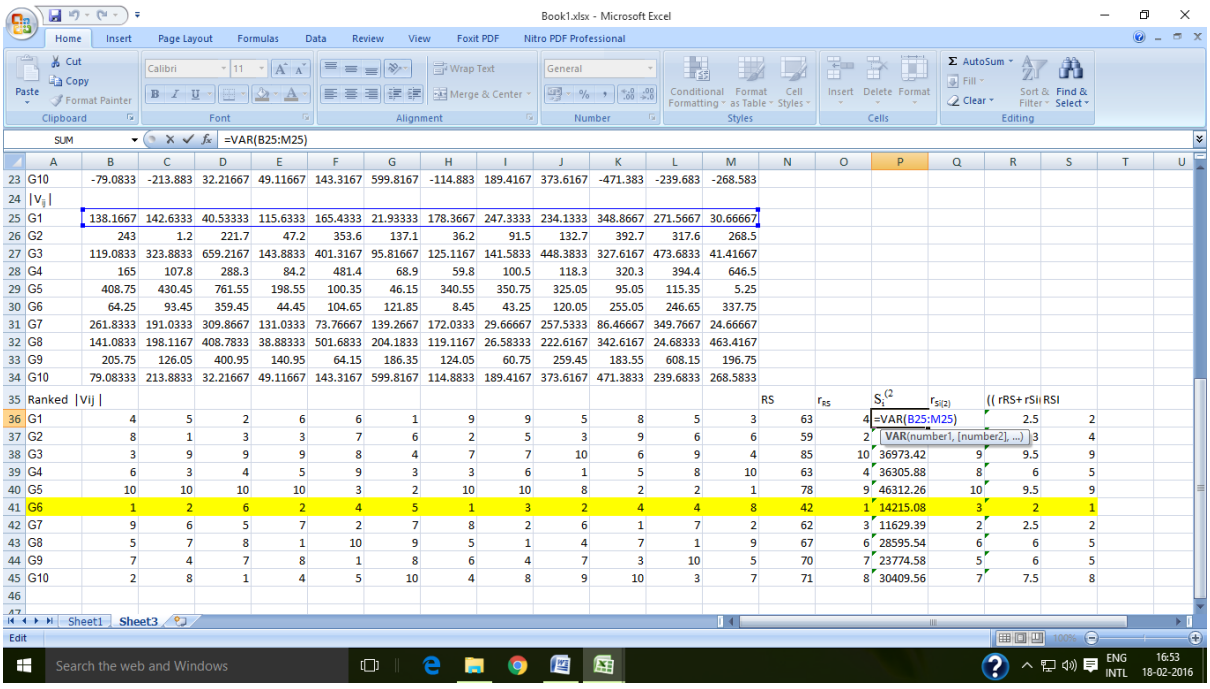
$$RSI = \text{rank of } ((r_{RS} + r_{Si(2)})/2),$$

उदाहरणतः

सबसे पहले हम ऊपर लिए गए डेटा से MS Excel में $V_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ का मान निकलतें हैंए उसके बाद $V_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ का निरपेक्ष मान निकालते हैं, प्रत्येक जीनोटाइप का सभी वातावरण में $V_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ का निरपेक्ष मान के आधार पर रैंकिंग करते हैं, उसके बाद $V_{ij} = (Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{Y}_{..})$ का निरपेक्ष मान के आधार पर प्रत्येक जीनोटाइप का सभी वातावरण में प्रसर.ा निकाल कर रैंकिंग करते हैं ।

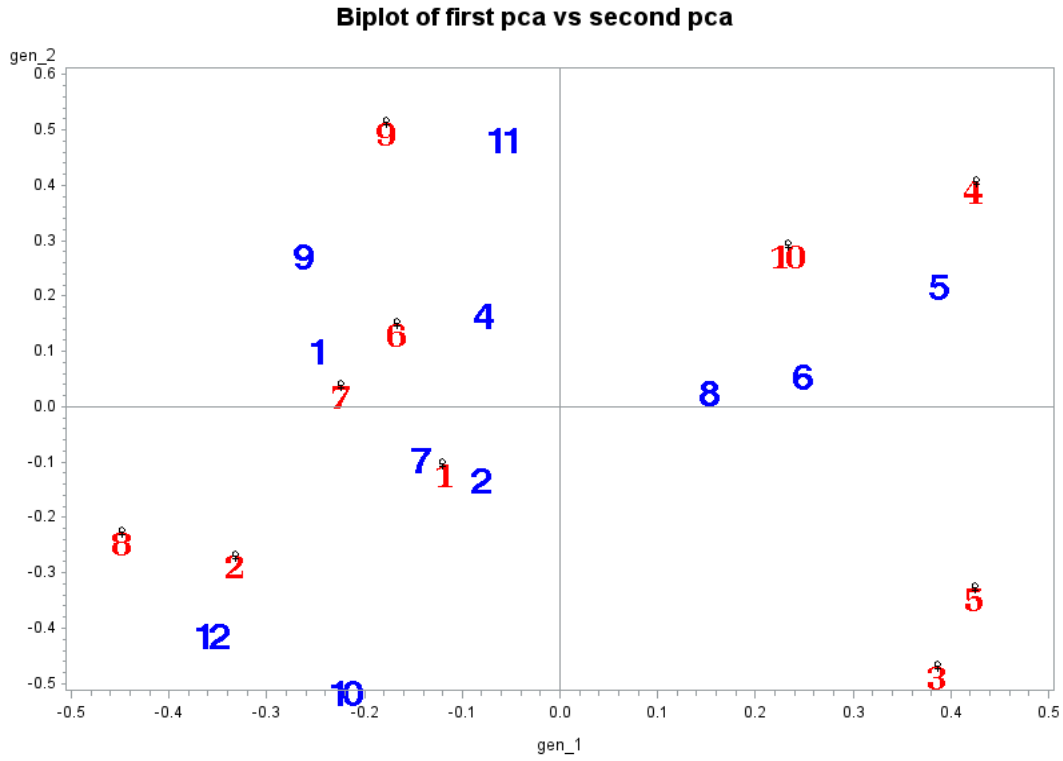


RSI= rank of ((r_{RS}+ $r_{Si(2)}$)/2)का मान निकालते हैं



निष्कर्षः प्रत्येक जीनोटाइप में से सबसे कम रैंक आधारित स्टेबिलिटी इंडेक्स (RSI)का मान वांछित परिणाम देता है ।

स्थायित्व के अप्राचालिक विधियों से प्राप्त परिणाम की तुलना के लिए हम उसी डाटासेट में additive main effects and multiplicative interaction (AMMI) model से biplotविश्लेषण से करने पर परिणाम संतोषप्रद मिलता है ।



कुछ चयनित संदर्भः

दावोलकर, ए. आर. (1999): "एलिमेंट्स ऑफ बायोमेट्रिकल जेनेटिक्स", कान्सेप्ट प्रकाशन कंपनी, नई दिल्ली, 373-374।

नास्सर, आर., एवं हुन, एम. (1987): "स्टडीज ओन एस्टीमेशन ऑफ फिनोटाइपिक स्टेबिलिटीरू टेस्ट ऑफ सिग्निफिकेंस ऑफ नॉन–पैरामीट्रिक मेसर्स ऑफ फिनोटिपिक स्टेबिलिटी", बॉयोमीट्रिक्स, 43, 45-53।

कांग, एम.एस. (1988): "ए रैंक–सम मेथड फॉर सेलेक्टिंग हाई इल्डिंग, स्टेबल कॉर्न जीनोटाइप", सिरियल रिसर्च कम्युनिकेशन, 16, 113-115।

राजू, बी. एम. के. एवं भाटिया, वि. के. (2003): "कोम्परिजन ऑफ वेरियस मेसार्स ऑफ स्टेबिलिटी विथ रेस्पेक्ट टू रैंकिंग एबिलिटी अंडर वर्यिंग सिचुएशन", जर्नल ऑफ इंडियन सोसाइटी ऑफ एग्रीकल्चरल स्टेटिस्टिक्स, 56 (3), 276-293 ।

शुक्ला जी. के. एच. (1972): "सम स्टैटिस्टिकल आस्पेक्ट ऑफ पार्टिशनिंग जीनोटाइप–एनवायरनमेंट कंपोनेंट्स ऑफ वरिअबिलिटी", हेरेडिटी, 29, 237-245 ।

राव, ,. आर. (1997): "जेनेटिक पैरामीटर्स के बेहतर आकलन के लिए कुछ योगदान", अप्रकाशित पी,चडी थीसिस, पी जी स्कूल, भा. कृ. अ. प., नई दिल्ली ।

रिक्की, जी. (1962): "क्षेत्र में अनुसंधान में जैविक विविधता को समझने की एक विधि", Z. Pflanzenzucht, 47, 92-46।

एबेर्हार्ट, एस.ए., और रसेल डब्लू,. (1966): "किस्मों की तुलना के लिए स्थायित्व मापदंडों", क्रॉप साइंस, 6ए 36-40 ।

गौच, एच.जी. जू., (1992): "क्षेत्रीय उपज परीक्षण के सांख्यिकीय विश्लेषण", एल्सेविएर, एम्सटर्डम ।

पर्किन्स, जे.एम. और विनोद, जे.एल. (1968). एन्विरोमेंटल और जीनोटाइप–एन्विरोमेंटल कंपोनेंट्स ऑफवरिअबिलिटी", III मल्टीप्ल लाइन्स एंड क्रोस्सेस. हेरेडिटी, 23,339-356 ।

जोबेल, आर.डब्लू., राइट, एम. जे. और गौच, एच.जी. जू., (1988): "उपज परीक्षण के सांख्यिकीय विश्लेषण", अग्रोनोमी, 80, 388-393 ।